Computational Linguistics and Chinese Language Processing
Vol. 12, No. 1, March 2007, pp. 1-16
© The Association for Computational Linguistics and Chinese Language Processing

Differences in the Speaking Styles of a Japanese Male According to Interlocutor; Showing the Effects of Affect in Conversational Speech

Nick Campbell[*+]

Abstract

There has been considerable interest recently in the processing of affect in spoken interactions. This paper presents an analysis of some conversational speech corpus data showing that the four prosodic characteristics, duration, pitch, power, and voicing all vary significantly according to both interlocutor differences and differences in familiarity over a fixed period of time with the same interlocutor.

Keywords: Conversational Speech Corpus, Expression of Affect, Prosodic Characteristics, Voice Quality Analysis.

1. Introduction

Human spoken interactions convey a variety of different types of information. In addition to the linguistic content of speech, there are also paralinguistic and extralinguistic elements that convey discourse-level and interpersonal levels of information related to the speaker, to the speaker's relationship(s) with the listener, and to the intended and actual progress of the discourse [Lindblom 1990; Stenström 1994; Hirschberg 1992, 1995].

Affect is conveyed in speech communication in a multitude of ways [Cahn 1989], including facial expression, gesture, body posture, speaking-style, tone-of-voice, lexical choice, syntactic construction, etc. It is perhaps impossible for a human to speak without revealing information about his or her affective states [Campbell 2005].

This paper examines how such affective information might be carried in the voice, particularly in the prosody of the speech, and shows from an examination of some corpus data

[*] National Institute of Information and Communications Technology
[+] Spoken Language Communication Research Laboratory, Advanced Telecommunications Research Institute International, Keihanna Science City, Kyoto 619-0288, Japan.
E-mail: nick@nict.go.jp; nick@atr.jp

[Received September 30, 2006; Revised October 30, 2006; Accepted November 1, 2006]

that evidence can be found for changes in affective state according to the nature of the interocutor and the history of their discoursal relationship.

2. The JST/CREST ESP Corpus

Speech research nowadays is predominantly corpus-based. One learns about the characteristics of speech and the expressivity of speech utterances from the analysis of a very large number of samples collected under a variety of speaking conditions [Campbell *et al*. 2006; Cowie *et al*. 2005]. For a period of five years, in order to aid the development of a technology capable of Expressive Speech Processing (ESP), the Japan Science and Technology Agency funded the collection of a large corpus of expressive speech that was coordinated by ATR in Kyoto, Japan.

As part of this corpus, over a period of three months during 2002, a group of ten volunteers were employed to talk with each other over the telephone for half-an-hour each time and to record their conversations to DAT using high-quality head-mounted condenser microphones. These conversations and their manually-produced transcriptions now form subset ESP_C of the JST/CREST ESP Corpus.

Of the ten volunteers, two were native speakers of Chinese, both fluent in Japanese, one male and one female, and two were native speakers of English, both fluent in Japanese, one male and one female. The remaining six were Japanese native speakers, three men and three women, living in the Kansai area of central Japan. They did not know each other initially but became familiar over the period of their telephone conversations. To our direct knowledge they never met face-to-face during this period.

This paper focuses on the speech characteristics of one male speaker from this corpus, JMA, who spoke with six partners over the three month period. In all, their conversations include 49,377 utterances from speaker JMA, where an utterance is approximately defined as the shortest meaningful unit of speech produced under a single intonation contour. The actual boundaries were determined on a case-by-case basis by the transcribers according to a set of rules published elsewhere [Campbell 2006].

The paper examines the acoustic characteristics of these utterances according to differences in interlocutor and stage of the interaction, showing that speaking style and voice phonation characteristics vary according to the interlocutor, in accordance with changes in familiarity and other speaker-listener relationships.

3. Materials for the Study

In previous work [Campbell 2005, 2006] affect-bearing utterances were distinguished from those that serve primarily to portray propositional or 'linguistic' content. The former, often called 'grunts' or 'affect bursts' are not usually found registered as words in a language dictionary, but are found very frequently in colloquial speech. For this study, a subset of 100 of those that occurred more than 50 times each in the conversations of one speaker (JMA) was selected, yielding 11,750 short conversational utterances for subsequent acoustic analysis. These were taken from five conversations each with each of the Chinese and English native-speakers, and from ten conversations with the Japanese native-speaker partners. Table 1 shows the number of utterances produced with each interlocutor. Table 2 lists the romanised orthographic transcriptions and counts of some of the more common examples.

Table 1. Utterance counts for the series of conversations with each interlocutor. The initial letters J,C,E in the interlocutor identifiers stand for Japanese, Chinese, and English respectively, the middle letters F and M stand for female and male respectively, and the third letter is an identifier.

CFA	CMA	EFA	EMA	JFA	JMB
1832	1632	1490	1773	2957	2066

Table 2. Counts of some of the more common utterance types that were studied for this paper, which occurred at least 50 times in the speech of JMA; the ten most common more than 1000 times. A sharp intake of breath was transcribed as '@S', a sniff as 'zu', and a laugh as '@W'. The letter 'n' indicates a syllabic nasal (umm). A minus sign represents moraic lengthening, which is known to be distinctive in Japanese, possibly triggering the percept of a different word. A dot represents a morphological boundary.

a 296	a- 368	a– 693	a— 608	a.a- 390	a-.hai 386
a.hai 577	a-.n 368	ano 337	ano- 494	a!! 927	demo 272
e- 665	e– 254	ee 2679	fun 642	fu-n 625	fu–n 273
ha.ai 978	hai 7295	ha-i 1657	hai.hai.hai 378	n(umm) 265	n- 456
n– 410	nanka 273	ne- 367	nee 284	@S 3382	sou 810
su- 429	su– 296	un 3717	u-n 2401	u–n 1243	u—n 333
un.un 351	@W 3041	zu- 1348	zu– 467		

The speech files corresponding to these utterances were analysed for their acoustic characteristics and a table of statistics for each utterance was produced. Specifically, the duration, pitch, power, and spectral characteristics of each utterance were recorded.

Duration was expressed both as absolute (log) duration of the measured utterance and as 'speaking rate' by dividing the absolute duration of the utterance by the number of phonemes in its transcription. This is a crude measure which does not take into consideration the inherent differences in different phone durations, but which serves to provide a simple approximation of speaking rate which will suffice for the present analysis.

Pitch (or more precisely, a measure of the fundamental frequency of the voice) was extracted using the ESPS 'get_f0' method that is incorporated in the 'Snack' signal processing library. The maximum and minimum pitch values for each file were recorded and stored along with an estimate of the range and average values for each utterance. The pitch contour was characterised by noting the average values measured over each third of the utterance, and stored these along with the percentage position of the pitch peak and the lowest pitch value.

Power values (*i.e,* measures of rms waveform amplitude) were calculated similarly, using the Snack command "power", and stored as maximum, minimum, average, and range for each utterance.

Spectral characteristics were calculated using the Snack command "dBPower" with options "-fftlen 128 -windowlength 128 -analysistype LPC -lpcorder 20". This produced an LPC-based 64-point vector representing the long-term average spectrum for the entire utterance (average length 0.54 seconds) from which values from points 2, 3, 4, 5, 7, and 9 were selected to represent the average power up to 1.5 kHz, points 12,15,19,23,28 to represent the average power between 1.5kHz and 4kHz, and points 34, 41, 49, 56, and 63 were selected to represent the average power between 4kHz and 8kHz. The average spectral energy measured in each of these three frequency bands was stored as a 3-valued vector for subsequent 'spectral' analysis. Since our main objective here is to examine spectral tilt, as evidence of differential phonation styles, the relative differences between the three bands (mid, high, and low on a mel-scale) were determined to suffice as a measure.

4. Analysis of the Data

After confirming independence of the variables under examination, a weak but insignificnt correlation of $r = 0.32$ was found between variations in pitch range and power range, and one of $r = 0.37$ between the averaged values of pitch and power across the 11,750 short utterances selected from the 49,377 utterances in the conversational corpus. There was a similar weak correlation between the measures of duration and power ($r = 0.34$) but none

between duration and pitch ($r = 0.19$). The correlation between spectral energy (power in the lowest band) and raw signal amplitude (signal power) was 0.08 . One can thus be satisfied that the measures are sufficiently independent to carry meaningful information in their differences.

There was a clear corrrelation of $r = 0.81$ observed between energy in the first spectral band (frequencies up to 1500 Hz) and in the second (frequencies between 1.5kHz and 4kHZ), but no such correlation between frequencies in the second and third bands (*i.e.,* between frequencies below and above 4kHz) which showed a correlation of $r = 0.2$. It is the difference between these latter two bands that is of interest here, since the lack of energy in the upper frequency bands is an indicator of a less tense, more breathy, speaking style which has been shown in previous studies ([Gauffin and Sundberg 1989; Sluijter and van Heuven 1994; Campbell and Mokhtari 2003]) to correlate with intimacy and a more careful manner of speaking.

4.1 Fundamental Frequency and Power

Figure 1 shows the values of f0 measured from the speech data of the male speaker JMA plotted separately for each interlocutor. The left plot shows average f0, the middle plot maximuim f0, and the right-hand plot minimum values of f0 measured in the conversations with each interlocutor respectively. The box-plots show median and interquartile values, with whiskers extending to 1.5 times the interquartile range. The boxes are drawn with widths proportional to the square-roots of the number of observations in the groups. A notch is drawn in each side of the boxes. If the notches of two plots do not overlap this is 'strong evidence' that the two medians differ at the 5% level of confidence.

Figure 1 shows that there is more variation in the voice fundamental frequency of speaker JMA when talking to the non-native partners, while the average values of f0 for the Japanese partners JFA and JMB are higher and less dispersed. The maximum f0 is highest when speaking with the English female, and lowest when talking with the Chinese male partner. When speaking with the Japanese native speakers, the maximum f0 shows the same median values as when talking with the English female partner, but there is overall more variety in f0 when speaking with the non-native parters.

Figure 2 plots the average, maximum and minimum power values for conversations with each of the six interlocutors. It shows that more energy is used when speaking with the Japanese partners, and more variation when speaking with the non-native interlocutors. Interestingly, the minimum power appears to be higher among conversations with the Chinese male partner than among conversations with the other interlocutors, but significantly lower for conversations with the Japanese female partner.

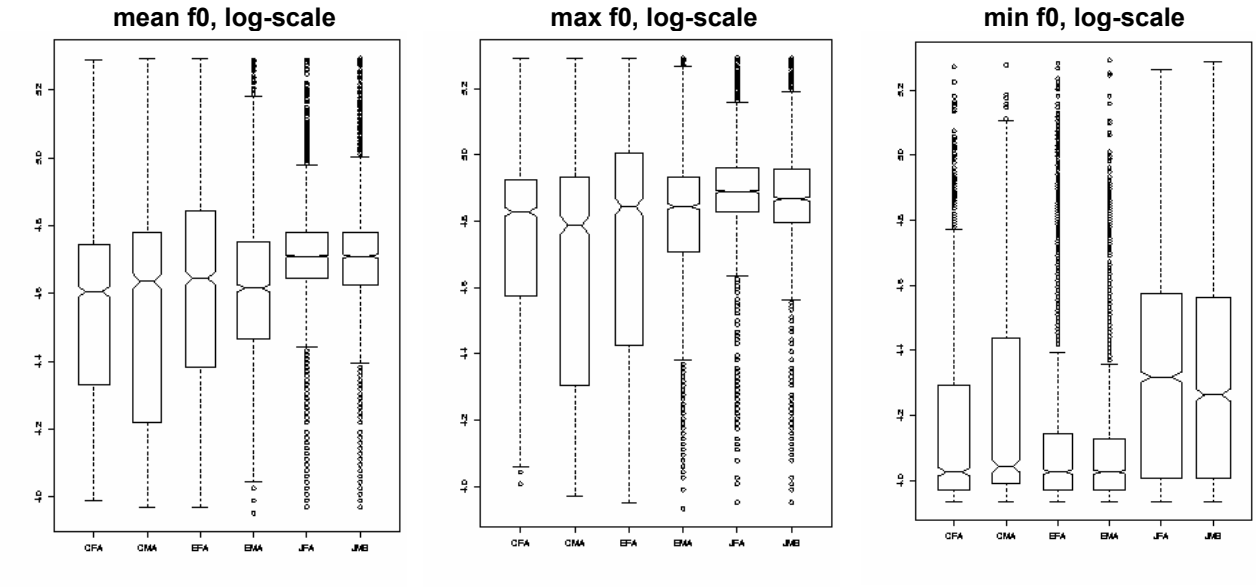


Figure 1. Plots of mean, maximum and minimum f0 values observed in the data of each of the interlocutors. The box-plots show median and interquartile values, with whiskers extending to 1.5 times the interquartile range. All F0 measurements are converted to their log values for ease of comparison.

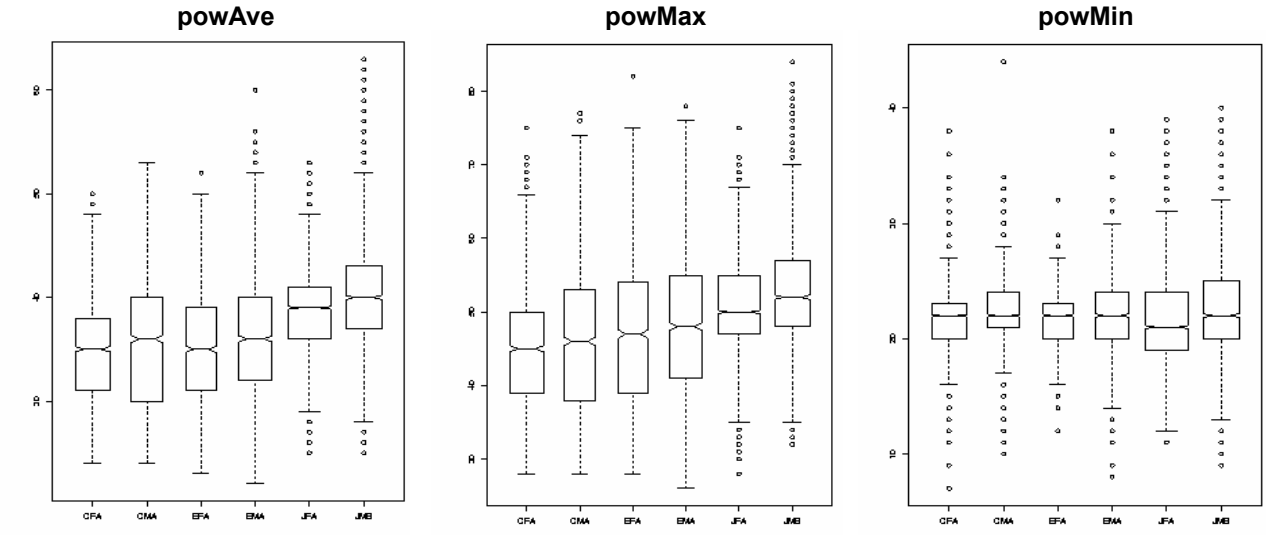


Figure 2. Plots of mean, maximum and minimum rms amplitude (speech signal power) values for each of the interlocutors.

Figure 3 plots f0 range for comparison with power range across the same set of partners. It shows a slightly higher range of f0 activity when this subject is talking with the English female than with the Chinese male partner. Both plots are log-scaled and show an average of

55Hz (*exp*(4)) median pitch range with a 30dB average power range when conversing with these different interlocutors. Power is noticeably higher when talking with Japanese native-speaker interlocutors, and lowest when talking with the female Chinese native-speaker. Needless to say, microphone distances (head-mounted) and record-level settings remained unchanged across all recordings.

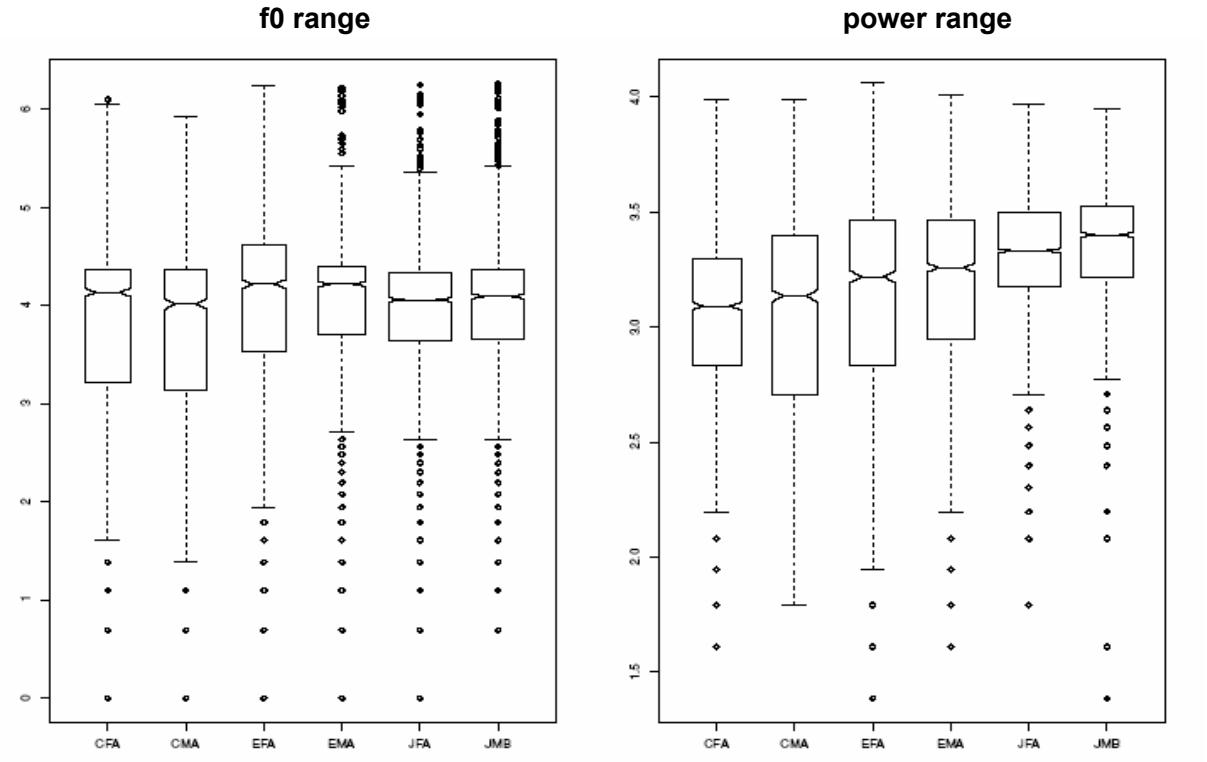


Figure 3. Ranges of fundamental frequency and power measurements for the affective utterances, plotted by interlocutor. As above, C, E, J stand for Chinese, English, Japanese respectively, and F, M represent female and male interlocutors. Both f0 and power are plotted as log values.

4.2 Duration and Speaking Rate

Figure 4 shows details of 'speaking rate' changes across the series of conversations with the two Japanese partners. This measure was calculated for each utterance by dividing the observed duration of its speech waveform (measured in milliseconds) by the number of characters in its transcription (see Table 2 for examples) and is therefore only an approximation of the true speaking rate, but it serves as a basis for comparison and provides a simple form of normalisation for the inherent differences in utterance type. Speaker JMA took part in nine conversations with female JFA, and eleven with male JMB. We note an average of 174.2 milliseconds per phone for interlocutor JFA, and an average of 169.65 for

interlocutor JMB. The speaking rate with the male partner appears to slow down throughout the series of conversations, while after reaching a peak in conversation J04 with JFA it appears to revert to a higher rate with the progression of time. Median values for JFA are 143, 162, 180, 183, 170, 166, 157, 171, and 158, while those for JMB are 144.5, 150.5, 183.0, 171.0, 149.5, 165.5, 170.5, 181.0, 167.5, 190.0, and 182.0.

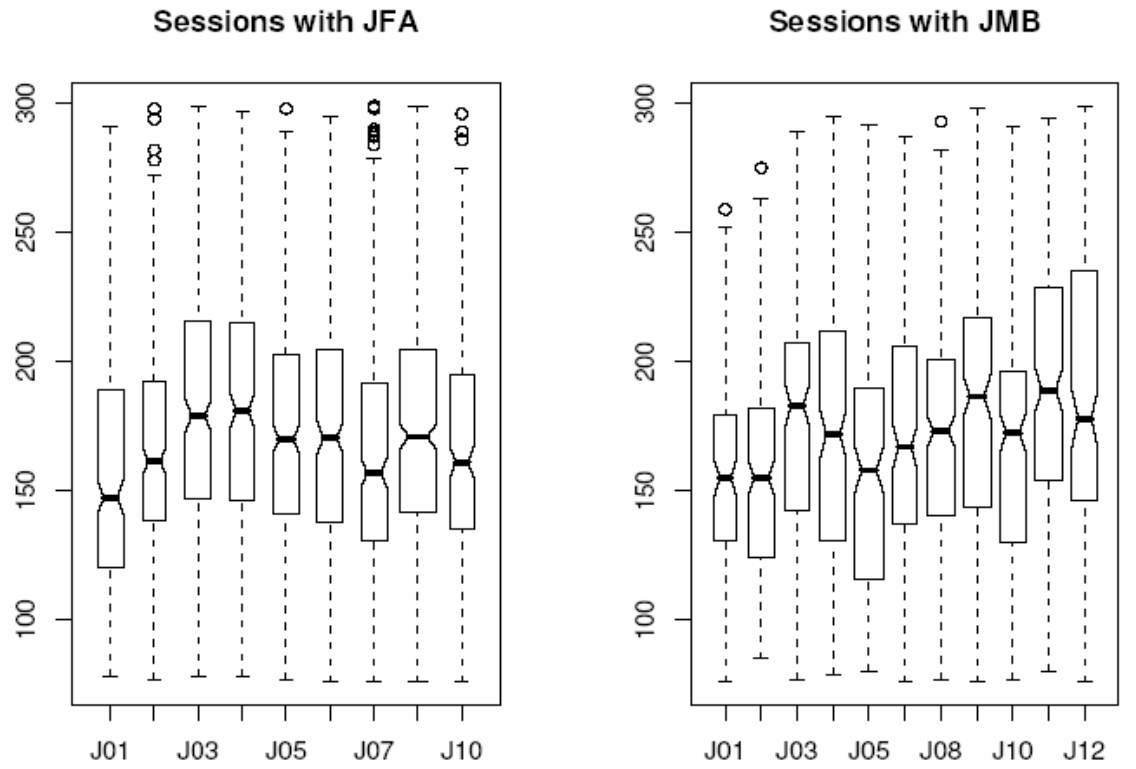


Figure 4. Speaking rate changes over weekly sessions

These values may seem unexpectedly long to an observant reader familiar with segmental durations, but it should be noted that they are sounds in afective grunts, not phones in lexical words. For example, the quantile durations (in seconds) for the word "hai" (="yes"/"I'm listening"/"I agree") are as follows: minimum: 0.152, 25th percentile: 0.382, median 0.43, 75th percentile: 0.49, and max: 1.59 seconds (n=7295). Those for the word "ee" (n=2679) are 0.268, 0.344, 0.42, 0.479, and 0.539. The durations observed for laughs in this context (n=3041) ranged from 119 milliseconds to four seconds, with a median duration of 0.9 seconds. Note that many of these utterances bear lengthening diacritics (e.g., 'a', 'a-', 'a–', 'a—', etc) and the transcribers who were all native speakers of Japanese were instructed to use one minus-sign to mark each mora-worth of lengthening perceived on the segment. It is customary to use such lengthening marks in standard Japanese Kana orthography, and mora durations are typically strictly observed in Japanese, where a moraic difference in timing

sructure can (unlike English) cue a different lexical item [1].

4.3 Spectral Slope

Spectral energy provides a simple cue to phonation style; the less energy in the higher part of the spectrum, the breathier the voice, and vice-versa. In pressed voice, the glottis closes faster as a proportion of the fundamental period, and the rapid closing is a sign of increased vocal effort and/or laryngeal muscular tension [Klasmeyer and Sendlmeier 1997; Fant 1993; Johnstone and Scherer 2000]. In conversational speech considerable use is made of voice phonation settings, especially for the display of affect [Campbell 2005; Campbell and Mokhtari 2003].

Figure 5 shows three measures of averaged spectral energy for the 11,750 affective utterances under examination. The low-frequency part of the spectrum is shown measured in decibels, as is customary in plots of spectral sections, but the middle and right-hand plots show differences between the low-frequency energy and the higher bands. Differences are also measured in decibels, and here 'low-band minus mid-band energy' is plotted in the centre plot, and 'mid-band minus high-band energy' is plotted in the right-hand plot. By plotting the differences rather than the absolute values, it is easier to visualise the spectral slope differences across these utterances.

The figure, averaged over all conversations, shows higher low-band energy for the Japanese female and the two Chinese interlocutors, with increased spectral slope for the Japanese female in the mid-band, and steeper spectral slope for the Japanese female and the two Chinese interlocutors at the top-end of the spectrum. The spectrum is therefore flatter overall for the English native-speaking partners and for the Japanese male partner. A flatter spectrum has been shown to reflect more tension in the voice.

Quantiles for the three spectral bands (measured over all data for speaker JMA) are given in Table 3, which shows median values to be -42, 14, and 9 decibels respectively. The difference of 14 decibels indicates that the average value for energy measured in the frequency range between 1.5kHz and 4kHz is -56 decibels, while the energy between 4kHz and 8kHz is typically at the -65 decibel level.

---

[1] For example, in Japanese, 'ie' means "house" while 'iie' with a longer first vowel means "no". Similarly, 'ka' is an interrogative particle, while 'ka-' with a lengthened vowel means "car". Such length-based lexical distinctions are common.

Table 3. Quantiles of the energy measured in three spectral bands. The top row shows absolute energy but the bottom two rows show energy differences measured in decibels.

	0%	25%	50%	75%	100%
low-band	-72	-48	-46	-44	-21
mid-band	-11	12	14	16	38
high-band	-18	7	9	11	37

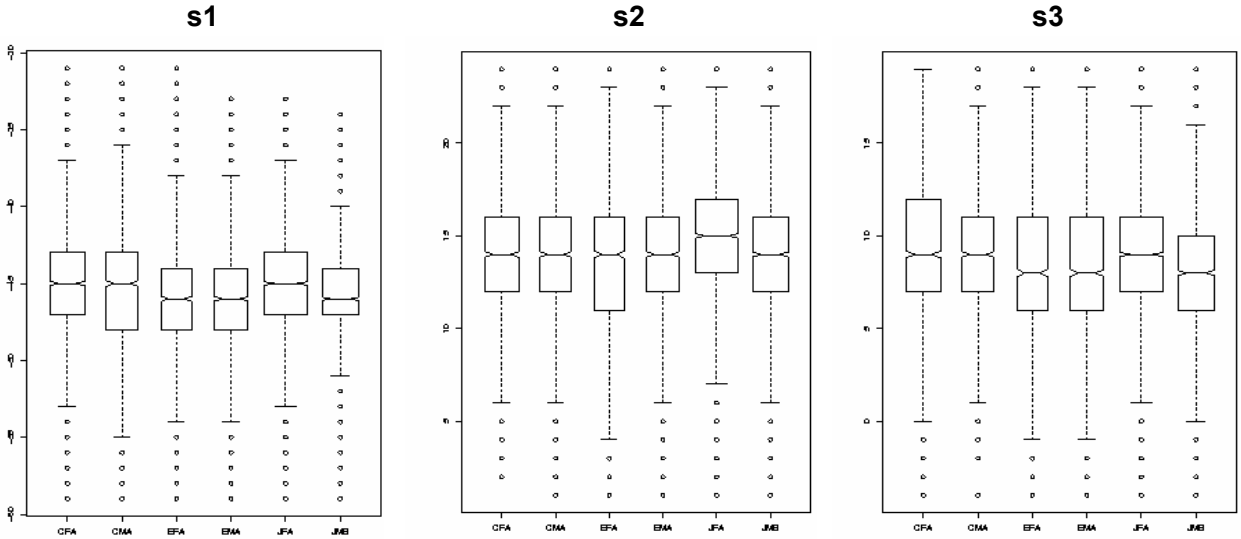


Figure 5. Three measures of spectral energy provide an indication of spectral slope. The left-hand plot shows average energy measured between 0-1.5kHz, the middle plot the difference between that and average energy measured between 1.5kHz and 4kHz, and the right-hand plot shows the difference between the averaged mid-band energy and the averaged energy between 4kHz and 8kHz at the top end of the spectrum. Measures are plotted separately by interlocutor.

Figure 6 shows differences in these values over time. The upper three plots show low-band, mid-band, and high-band energy measures for conversations with Japanese male partner JMB. The lower part of the figure shows only the high-band energy differences (spectral slope measures) for the four non-native partners.

In each case there is a general trend towards decrease in steepness of the spectral slope with time. From the top plots (of the series of conversations with partner JMB), the second, third and penultimate conversation exhibited high low-frequency spectral energy (from the left-hand plot), steep falloff in mid-band energy (from the central plot), and, at least for the

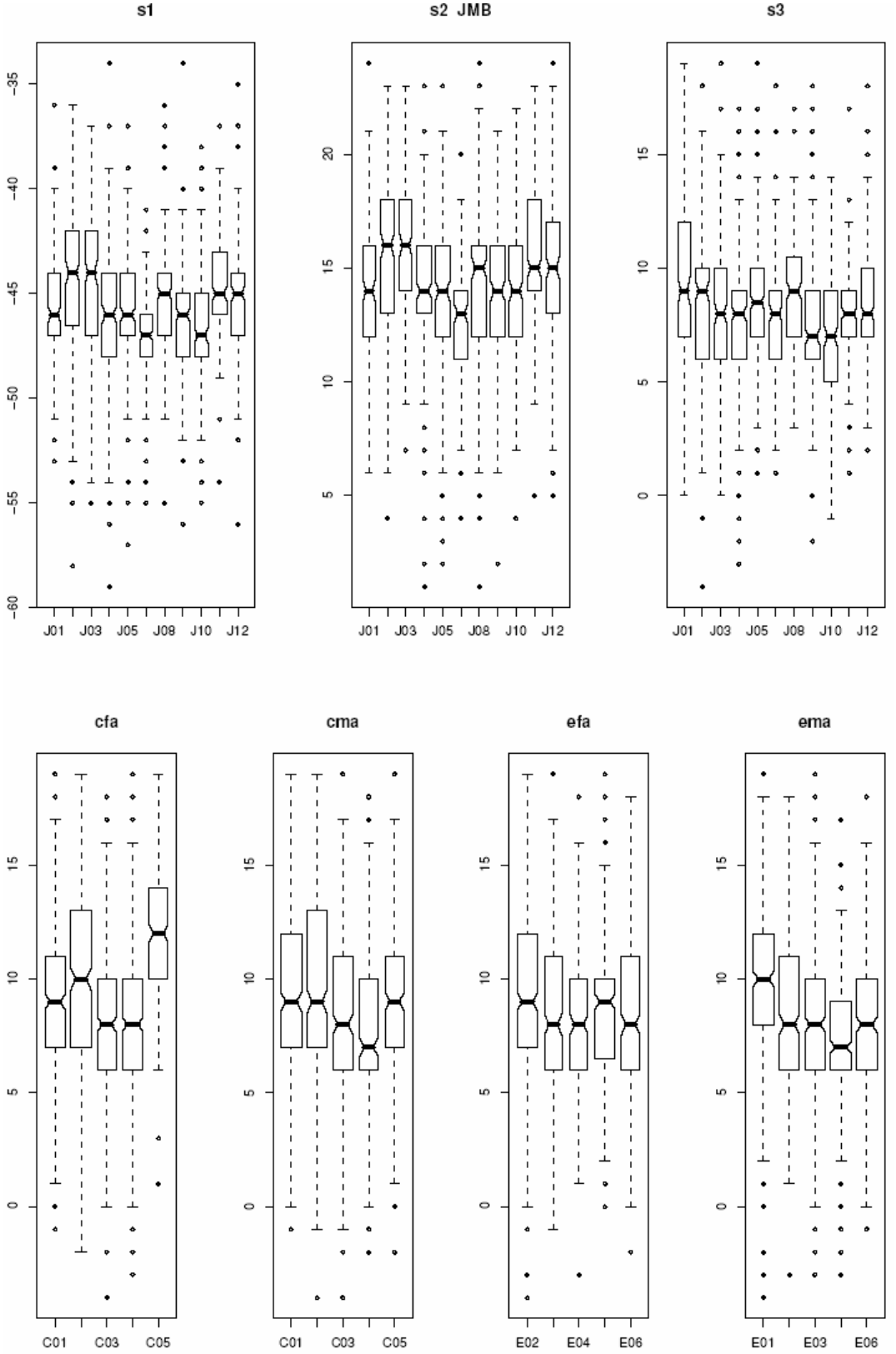


Figure 6. Spectral slope differences across time by interlocutor. The top part shows all three spectral bands for partner JMB, and the bottom part shows the difference between mid-band and high-band energy for each of the non-native partners.

second and third conversations, considerable variability in the high-frequency dropoff. This would be consistent with a higher degree of tension and varying politeness in the speech of the initial and penultimate conversations. For conversations in the interim period, from J04 to J10, a gradual decrease of steepness is found in the high-end spectral tilt that would be consistent with an increase in familiarity as reflected by more frankness and less polite softening of the voice. Then for the final conversations, as the recordings (and the three-month relationship) come to an end, there is an increase again, as would be consistent with a rise in formality of the conversational speech between the partners.

In the lower plots, with the non-native speaker partners, there is a similar steeepness in high-frequency dropoff (consistent with increased politeness in the voice and speaking style) in the initial and final recordings, and a gradual relaxation of spectral tilt in conversations of the interim period. Steeper spectral slope is found in the conversations with the female partners, with the Chinese female being highest and the English male being lowest in this respect.

5. Discussion

In this analysis of the prosodic characteristics of the conversational speech of one Japanese male over a period of three months, considerable variation was found in all of the parameters measured. By factoring the analysis according to differences in interlocutor as well as by differences in time, or sequence of the conversations, we were able to show that the changes are not a result of time-related changes, such as tiredness or ill-health, but that they correlate more with differences in interlocutor and with development of the individual relationships.

It is probable that not all interlocutors were related to in an equal way. One can imagine more sharing of common interests between native speakers of the same language, and different forms of bonding in the relationships that developed between the male and the female partners respectively. Similarly, the culturally closer, Asian but foreign, Chinese partners and the possibly exotic, and maybe more foreign, English speakers would have brought different contributions and cultural assumptions to the conversations. Their necessary lack of fluency in the use of Japanese, particularly over the telephone where the visual support for communication is impaired, would have introduced idiosyncracies into the style of the different conversations.

Without a complementary analysis of the texts of the conversations, one can only draw speculative conclusions to explain the differences in the prosodic characteristics, but from the spectra of speech with partner JMB in Figure 6 it can be assumed that the initial relatively low spectral energy and high spectral tilt of conversation J01 represent the ‘baseline’ settings for speaker JMA who had no expectations at that time about his partners. One might then

speculate that the apparent increase in politeness (as indicated by a more breathy speaking style) in conversations II and III could be due to having to maintain a conversation for a long 30-minutes over the telephone with a partner who is still relatively unknown to the speaker, and that the decrease thereafter occurred as they found more interests in common to talk about. From a brief examination of the transcriptions, they certainly appear to have become friends over the three month period. If so, then perhaps one can also speculate that the increase of breathiness in their speech towards the end is indeed due to the approaching termination of their telephone relationship.

6. Conclusion

In light of the recent considerable interest in the processing of affect in spoken interactions, an analysis was performed of some corpus data of conversational speech, showing that the four prosodic characteristics, duration, pitch, power, and voicing all vary significantly according to interlocutor differences and to differences in familiarity and politeness over a fixed period of time with the same interlocutor.

The results showed significant differences in the prosodic characteristics of speech with others sharing the same native language as compared with those of non-native speakers of Japanese. The results also showed that speaking rate, pitch range, and spectral tilt varied significantly according to partner and position of the conversation in the three-month series. Because different settings were used with different partners at the same time, the possibility can be discounted that these differences were due to unrelated external considerations such as variation in the health of the speaker.

The findings reported earlier for similar changes in phonation settings for a female speaker from a separate section of the corpus (see [Campbell 2005; Campbell and Mokhtari 2003]) under more varied conversational settings have been replicated here with data from a different speaker in a more controlled recording environment.

It is perhaps still too early to make use of these findings in speech technology, and considerable further work is required before strong claims can be made about the causes and relationships, but it is of interest that these differences exist at all. Listeners certainly make use of small but consistent speaking-style and phonation-setting changes to make inferences about the affective states of the speaker. Perhaps these variations will provide the foundation for both speech synthesis and speech recognition modules that begin to incorporate affect as one of the strands of meaning in speech. Such technology would be of great use in providing a softer interface between machines and humans in society.

References

Cahn, J., "The generation of affect in synthesised speech," *Journal of the American Voice I/O Society,* Vol c8, 1989, pp. 251-256.

Campbell, N., "Getting to the heart of the matter; speech as expression of affect rather than just text or language," *Language Resources & Evaluation,* 39(1), Springer, 2005, pp. 109-118.

Campbell, N., "Conversational Speech Synthesis and the Need for Some Laughter," *IEEE Transactions on Audio, Speech, and Language Processing,* 14(4), July 2006.

Campbell, N., "A Language-Resources Approach to Emotion: Corpora for the Analysis of Expressive Speech," In *Proc International Conference on Language Resources and Evaluation*, LREC 2006.

Campbell, N., L. Devillers, E. Douglas-Cowie, V. Auberge, A. Batliner, and J. Tao, "Resources for the Processing of Affect in Interactions," Panel session, In *Proc LREC'06*, Genoa, Italy, 2006, pp. xxiv-xxvii.

Campbell, N., and P. Mokhtari, "Voice Quality; the 4th prosodic parameter," In *Proc 15th ICPhS*, Barcelona, Spain, 2003.

Cowie, R., E. Douglas-Cowie, and C. Cox, "Beyond emotion archetypes; Databases for emotion modelling using neural networks," In *Neural Networks 18*, 2005, pp. 371-388.

Fant, G., "Some problems in voice source analysis," *Speech Communication,* 13(1), 1993, pp. 7-22.

Gauffin, J., and J. Sundberg, ''Spectral correlates of glottal voice source waveform characteristics," *Journal of Speech and Hearing Research,* 32, 1989, pp. 556-565.

Hirschberg, J., ''Using discourse content to guide pitch accent decisions in synthetic speech," In G. Bailly and C. Benoit, ed, *Talking Machines*, North-Holland, 1992, pp. 367–376.

Hirschberg, J., ''Acoustic and prosodic cues to speaking style in spontaneous and read speech," In *Symposium on speaking styles, Proc ICPhS*, Stockholm, Sweden. 1995.

Johnstone, T. and K. R. Scherer, "Vocal Communication of Emotion," in: M. Lewis & J. Haviland (Eds.) Handbook of Emotion (2nd ed.). New York: Guildford. 2000.

Klasmeyer, G., and W. F. Sendlmeier, "The classification of different phonation types in emotional and neutral speech," *Forensic Linguistics,* 4(1), 1997, pp. 104-124.

Lindblom, B. E. F., ''Explaining phonetic variation: A sketch of the H&H theory," In *Speech Producstion and Speech Modelling, NATO-ASI Series D: Behavioural and Social Sciences,* edited by H. J. Hardcastle and A. Marchal (Kluwer, Dordrecht), Vol 55, 1990.

Schroeder, M., "Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions," In *Proc. Workshop on Affective Dialogue Systems: Lecture Notes in Computer Science*, Kloster Irsee, Germany, 2004, pp. 209-220.

Sluijter, A. M. C., and V. J. van Heuven, "Spectral tilt as a clue for linguistic stress," presented at 127th ASA, Cambridge, MA. 1994.

Stenström, A., *An Introduction to Spoken Interaction*. Longman, London. 1994.

Website Resources

Snack: a Tcl/Tk library and toolkit for speech signal processing. http://www.speech.kth.se

The Japan Science & Technology Agency *Core Research for Evolutional Science & Technology*, 2000-2005.

The JST/CREST Expressive Speech Processing Project homepage can be found at http://feast.atr.jp/

Computational Linguistics and Chinese Language Processing
Vol. 12, No. 1, March 2007, pp. 17-32
© The Association for Computational Linguistics and Chinese Language Processing

The Breath Segment in Expressive Speech

Chu Yuan*, and Aijun Li*

Abstract

This paper, based on a selected one hour of expressive speech, is a pilot study on how to use breath segments to get more natural and expressive speech. It mainly deals with the status of when the breath segments occur and how the acoustic features are affected by the speaker's emotional states in terms of valence and activation. Statistical analysis is made to investigate the relationship between the length and intensity of the breath segments and the two state parameters. Finally, a perceptual experiment is conducted by employing the analysis results to synthesized speech, the results of which demonstrate that breath segment insertion can help improve the expressiveness and naturalness of the synthesized speech.

Keywords: Breath Segment, Expressive Speech, Emotion, Valence, Activation

1. Introduction

In the current speech synthesis and recognition systems, some characteristics of spontaneous speech are treated as noise, such as disfluent utterances, repeated sounds, filled pauses, salient breaths and coughs. In corpus collection for speech synthesis and recognition systems, the speaking style of the speakers is always strictly controlled and the speaker is usually required to give a "canonical pronunciation" to decrease the speaking noise as much as possible. However, in recent study, researchers have begun to pay more attention to the non-verbal information in natural speech, especially the paralinguistic and physiological information. They have focused on how to use these types of information to improve the naturalness and expressiveness of emotion and attitude in synthesized speech, so that the speaker's intention can be better understood during verbal communication.

In 1989, Cahn compiled a simple feeling editor based on the phonetic characteristics of emotion [Cahn 1990]. Vroomen, Collier and Mozziconacci examined the duration and intonation of emotional speech and proposed that emotions can be expressed accurately by manipulating pitch and duration based on rules. This conclusion showed that, in emotional

* Institute of Linguistics, Chinese Academy of Social Sciences, No. 5 Jianguomennei Dajie, Beijing, 100732 China
E-mail: Yuanchu8341@gmail.com; liaj@cass.org.cn

[Received August 12, 2006; Revised October 22, 2006; Accepted November 1, 2006]

speech, duration and intonation can be employed to observe the speakers' attitude [Vroomen *et al*. 1993]. In 1998, Campbell found that if one compares the same content in different forms, for example, a news item in its read form, its formal spoken or broadcast form, and its informal conversational form, differences are obvious not only in lexis, word-order, chunking, and prominence relations, but also in the mood of the speaker and in the tone of the voice [Campbell 1998].

In 2000, the International Workshop on Speech and Emotion of ISCA (held in Ireland) invited, for the first time, researchers who were devoted to the study of emotion and speech. Before this conference, many researchers had begun to investigate the voice quality, prosodic features, and acoustic features of emotional speech. Alku and Vilkman designed an experiment to illustrate that the phonation types could be separated from each other effectively when the quantification was based on the parameters extracted from the instant of the maximal glottal opening and the minimal peak of the flow derivative [Alku *et al*. 1996]. Heuft, Portele, and Rauth carried out a more sophisticated test in order to determine the influence of the prosodic parameters in the perception of a speaker's emotional state in three different testing procedures. Their studies proved that the recognition rates were lower than those in the preliminary test, although the differences between the recognition rates of natural vs. synthetic speech were comparable in both tests. The outcome of the saw tooth test showed that the amount of information about the speaker's emotional state transported by F0, energy, and overall duration was rather small. However, the relations between the acoustic, prosodic parameters, and the emotional content of speech could be determined [Heuft *et al.* 1996]. Iida recorded a corpus of one speaker which included three kinds of emotion: anger, happiness and sadness. When synthesizing emotional speech, they picked up the corresponding emotional segments from the emotion corpus. The emotion speech, synthesized in this way, achieved a correct recognition rate 50% ~80% higher than through previous means [Iida *et al*. 2000]. Campbell focused on how to express a modal word in spontaneous speech with various emotions and attitudes [Campbell 2004].

Some researchers have also studied the non-verbal information in emotional speech. Trouvain attempted to analyze the terminological variety from a phonetic perspective. He proposed that the overview of various types of laughter indicated that further concepts of description were needed. In a pilot study on a small corpus of spontaneous laughter, the usefulness of the concepts and terms in practice was examined [Trouvain 2003].

In the light of the above overview of emotion speech research, this paper mainly discusses the function of the non-verbal information in natural speech, specifically the common non-verbal information which includes breath, laugh, filled pause, long silence, and cry. The breath segment is taken as an example to observe how the acoustic characteristics are related to prosodic structure, expressive valence, and activation through statistic analysis of

reading and spontaneous speech. The concluded rules are then applied to a perceptual experiment to see how it works.

2. Materials

2.1 Breath Segments

This paper studies breath segments which appear in both read and spontaneous speech, as shown in Figures 1 and 2, annotated between two dotted lines in the read and spontaneous speech, respectively.

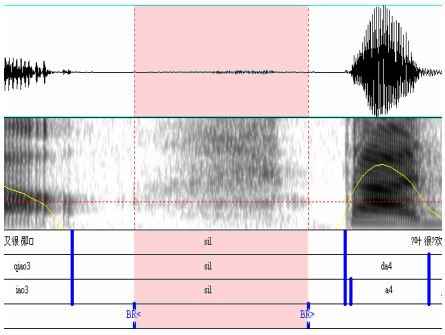


Figure 1. Breath segment in reading speech

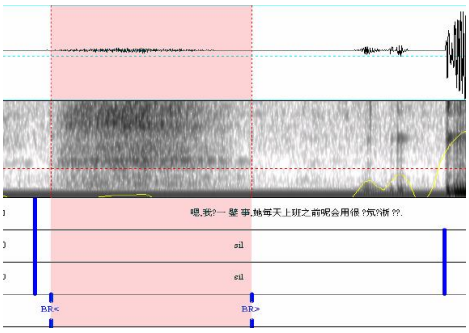


Figure 2. Breath segment in spontaneous speech

The breath shown here is not the normal unconscious physiological exhalation or inspiration process but the deliberate breath for expressing a kind of emotion. Therefore, the following breath segment carries the emotional or attitudinal information of the utterance. Moreover, the acoustic features, such as the length and intensity of the breath segment, may be correlated to the emotional state in terms of valence and activation. Further, the small blanks preceding and following the breath segment which are caused by the physiological need of a breath segment may be inserted when the synthesis of emotional speech is conducted.

The breath has two functions: fulfilling the physiological requirement of the intake of air and the expression of emotion or attitude. The authors determine the activation and valence degrees for each recitation of each phrase and use the information to label the breath segment before this phrase.

2.2 The Corpus and Annotation

The corpus used in this paper is called CASS-EXP which includes read and spontaneous speech. The first part contains some stories read by actors and actresses in emotional and neutral states while the second part includes TV and radio programs along with spontaneous speech: monologues and dialogues.

SAMPA-C [Li 2002] and C-ToBI [Chen *et al.* 2000] are adopted to label segmental and prosodic information. Furthermore, the starting and ending points of breath segments in terms of valence and activation degrees are labeled as well.

The authors labeled the emotion characteristics of the breath segments based on two factors: valence and activation. The theoretical foundation of valence is the concept of a separation of positive or negative emotion. The function of activation is the enabled degree of energy which is in contact with the emotion condition. The activation and valence of one breath segment here refer to the activation and valence of the following intonational phrase.

Emotional valence is categorized into three levels: positive (1), neutral (0) and negative (-1). The activation has three categories as well: excited (1), steady (0) and low (0). When both the emotional valence and activation of a certain breath segment are marked as 0, the breath segment is considered to be a neutral physiological segment without carrying any expressive information.

Three boundary levels (break index) 1, 2, 3 are annotated which stand for prosodic word, minor prosodic phrase, and major prosodic phrase (intonational phrase), respectively. The authors intend to examine whether the breath segment occurs in a normal stop or in an unexpected position. The normal stop refers to the breath at a prosodic phrase boundary, and the unexpected or abnormal position is the breath at a prosodic word boundary or within a prosodic word.

3. Breath Segments in Read Speech

From CASS-EXP, the authors select fifteen fragments from a read story which have different emotional states and attitudes. The valence and activity of nine fragments were labeled.

3.1 Occurring Number and Position of the Breath Segments

Based on what has been labeled, the number of breath segments is calculated for neutral and expressive speech. It was found that the number of breath segments in expressive speech is 50% higher than in that of neutral read speech in the same text. In these nine fragments, the number of breath segments in expressive speech is 334, and only one of them appears in an abnormal stop; the number in neutral speech is 225, of which all appear in normal boundaries, as shown in figure 3.

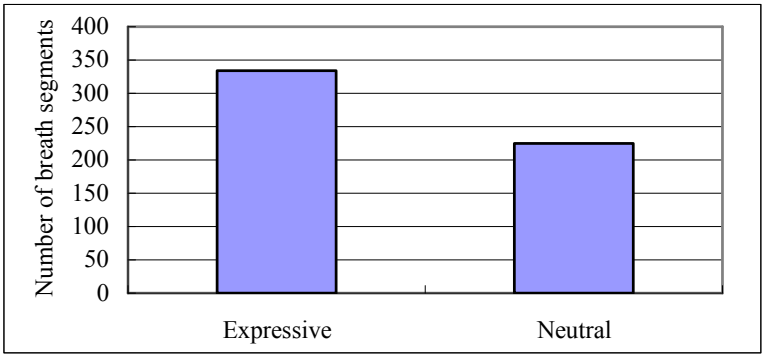


Figure 3. Number of breath segments in expressive and neutral read speech.

In fragments of read form, most of the breath segments occur at boundary 3 (intonational phrase boundary). The number of the breath segments at boundary 1 (prosodic word boundary) is the smallest, as shown in Figure 4. Table 1 demonstrates that the boundary distribution of breath segments appearing in expressive speech and neutral speech exhibits no difference. In expressive and neutral speech, the number of breath segments at boundary 1 is the smallest, and the number of breath segments at boundary 3 is the largest.

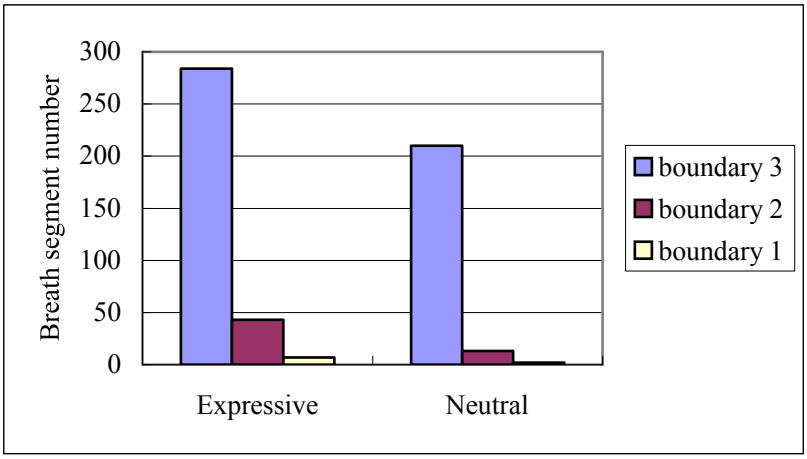


Figure 4. The number of breath segments at the different boundaries.

Table 1. Number and percent of breath segments of emotion and neutral read speech at the different boundaries.

Boundary	Number of breath segments in expressive speech	Percent	Number of breath segments in neutral speech	Percent
3	284	85.2%	210	93.3%
2	43	12.8%	13	5.8%
1	7	2%	2	0.9%

In general, breath segments in read speech, either expressive or neutral, usually appear between two prosodic phrases, especially between two intonational phrases. From the perspective of syntactic analysis, most of the breath segments appear between two intonational phrases or two intonational phrase groups.

We measured the duration of the silence which was between the breath segment and the prosodic phrase following this breath segment. The mean duration of the silence in different valence and activity is shown in Table 2.

Table 2 .The mean duration of the silence in different valence and activity

	Valence			Activity		
	-1	0	1	-1	0	1
Emotional	64ms	54ms	40ms	78ms	52ms	28ms
Neutral		48ms			49ms	

From this table we can know that in neutral speech the duration of the silence which was between the breath segment simply and the prosodic phrase following this breath segment is about 50ms. In emotional speech the durations are different because of the different valence and activity.

3.2 Duration of Breath Segments in Read Speech

In these nine fragments whose valence and activity have been labeled, the number of breath segments in expressive speech is 200, and only one of them appears in an abnormal stop; the number in neutral speech is 133, of which all appear in normal boundaries.

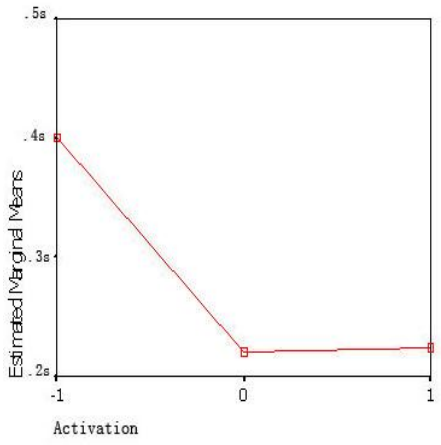


Figure 5. Breath segment mean duration and activation

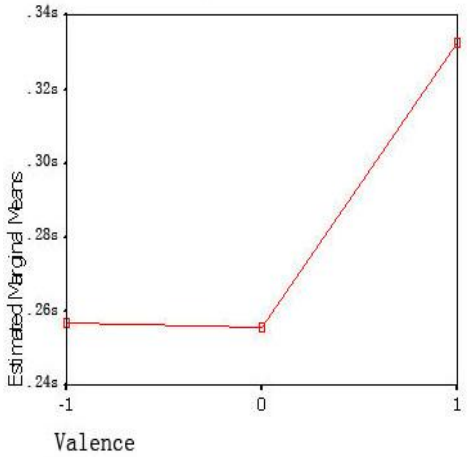


Figure 6. Breath segment mean duration and valence

The durations of breath segments are measured and put into a multi-variance analysis using SPSS. Breath segment means are shown in figure 5 and figure 6. In the analysis of the relationship between the valence degree and the duration of the breath segment, it was found that there is no significant correlation between the three categories of emotion valence and the duration of the breath segment (P=0.063>0.05).

However, activation has significant influence on the breath duration (P=0.000<0.05). The result of the analysis indicates that when the activation is 0 or 1, the discriminative degree of duration is not very high; when the activation is -1, the degree is different from that in other two activation states.

Table 3. Tests of between-subjects: valence and activation effects to the duration and intensity of breath segment.

Source	Dependent Variable	F	Sig.
valence	intensity	.544	.581
	duration	2.801	.063
activation	intensity	10.313	.000
	duration	9.344	.000
valence* activation	intensity	.371	.829
	duration	2.092	.083

Table 3 displays the effect triggered by valence and activation on intensity and duration. The valence has no effect on the breath duration and there is no interactive effect of valence and activation on intensity and little on duration (P=0.083). This result proves that, although the speakers express a certain kind of emotion, the physiological response does not differ from that of neutral speech. Nevertheless, because we do not know that the compute method in SPSS is the same as the person's mental perception mechanism or not. In this kind of case, we think that the effect triggered by valence and activation has influence of breath segments.

In addition to the duration of breath segments, the authors computed the intervals between two breath segments and their distribution. Among the 319 intervals there were 304 intervals shorter than 10 seconds. The other 15 intervals which include error reading were longer than 10 seconds. So this confirms that, when a text is read at normal speed, the time between two breath segments is shorter than 10 seconds.

3.3 Intensity of Breath Segments

Another important characteristic is the intensity of the breath segments. Tables 4 and 5 are the statistical results on intensity grouped by valence and activation.

Table 4. Breath segment intensity grouped by valence

Valence	N	Subset	
		1	2
0	155	37.8143	
1	29		41.9793
-1	16		43.8315
Sig.		1.000	.202

Table 5. Breath segment intensity grouped by activation

Activation	N	Subset		
		1	2	3
0	120	36.5159		
-1	21		39.5437	
1	59			43.5185
Sig.		1.000	1.000	1.000

Afterwards, the authors observed the relationship between the intensity of every breath segment and the intensity of the following intonational phrase. Through the examination of the data obtained from SPSS analysis which be shown in table 6, it was found that activation has a significant effect on the intensity ratio of the following intonational phrase in the breath segment; in addition, the effect of valence and the interactive effect of valence and activation are significant as well.

Table 6. Tests of between-subjects effects which is valence and activation effects to the IR

Source	Sig.
Activation	.022
Valence	.913
Activation * Valence	.609

Table 7 provides the means and ranges of intensity ratios of the following intonational phrase to the present breath segment (IR) in three categories of activations. The intensity ratio is the lowest when the activation is 0.

Table 7. The means and ranges of intensity ratios in three categories of activation

Activation	Mean	95% Confidence Interval	
		Upper Bound	Lower Bound
-1.00	0.634	0.682	0.592
0.00	0.558	0.573	0.544
1.00	0.646	0.674	0.619

3.4 Rules of Inserting Breath Segments to Read Speech

One can obtain rules of breath segment insertion based on the previous analysis of synthesized speech. The breath segment corpus can be set up first for the selected speaker. When the speech is being synthesized, the fitted breath segments can be selected and inserted into the expected positions. The insertion rules are summarized as follows:

A. At every major prosodic phrase boundary, a breath segment can be inserted or produced. The durations of these breath segments are about 0.5 second or longer.

B. Intervals between two breath segments are no longer than 10 seconds, *i.e.* one sentence group length in text is shorter than 10 seconds.

C. Within one intonational group, the number of the breath segments is uncertain, generally, there are one or two breath segments before longer intonational phrase and the breath duration ranges from 0.1 to 0.3 second.

D. When the activation of breath segment is not 0, the intensity of this breath segment is set to 0.6 -0.7 times of the intensity of following prosodic phrase. When the activation of breath segment is 0, the intensity of this breath segment is 0.5 times of the intensity of the following prosodic phrase.

E. Between every breath segment and the prosodic phrase following this breath segment there is a silence.

F. The duration range of different kind of valence and activation is induced from the read speech. The breath segment in the synthesized speech is selected random in the range of corresponding kind.

Although the breath segment is not the only way to express emotion or attitude in read speech, breath segments inserted in the synthetic speech can prompt the naturalness and expressiveness. Also, the synthesis speech with breathy segment insertion is more acceptable to the subjects.

4. Breath Segments in Spontaneous Speech

The authors select nine dialogs from the CASS-EXP corpus. Each dialog is a conversation between an audience and a radio host through a hotline telephone. It is assumed that the radio hostess's emotion is the performed emotion while the audience's is natural. In this part, boundary 4 is used to label the turn taking boundary.

4.1 Positions of Breath Segments in Spontaneous Speech

In these nine dialogs, 55 breath segments produced by the radio hostess and 17 breath segments are at abnormal positions, *i.e*. unexpected prosodic boundaries, which account for about 32% of the total breath segments. The audiences make, altogether, 54 noticeable breaths

at normal boundaries and 19 at abnormal ones ，which occupy about 35.2% of the total.

The radio hostess produces 11 physiological breath segments while the audience produces only 6. These 17 segments all appear at major prosodic phrase boundaries. In general, the physiological breaths that appear in spontaneous speech are similar with those in read speech but the frequency of appearance declines greatly.

From Table 8, one can see that the distribution of the physiological breath segments produced by the radio hostess is well-proportioned. The physiological breath segments produced by the audiences appear at boundaries 3 (prosodic phrase) or 4 (turn taking). Thus, the data help prove that when the expressiveness is a performed one, the breath distribution is the same as that in neutral speech. However, for spontaneous speech with natural expression (in Table 9), the breath also appears at boundaries 1 and 2. So, one can confirm that, in natural emotion speech, most of boundaries 1 and 2 are made intentionally. If one synthesizes this kind of speech material, one can consider breaking the original prosodic structures by adding breath segments.

Table 8. The breath segment distribution at prosodic boundaries by the radio hostess

Boundary	Total	Abnormal position	Normal position	Physiological breath
1	6	6	0	2
2	23	10	13	4
3	16	1	15	3
4	10	2	8	2

Table 9. The breath segment distribution at prosodic boundaries by the audiences

Boundary	Total	Abnormal position	Normal position	Physiological breath
1	9	6	3	0
2	9	8	1	0
3	14	2	12	2
4	22	3	19	4

4.2 Duration of Breath Segments in Spontaneous Speech

Figures 7 and 8 show the duration distribution of the breath segments made by the radio hostess according to valence and activation. The bottom and top value are 25% and 75% accumulative frequency, respectively, standing for duration variation range. (Note that in Figure 7, when activation is -1, the token number is relative small).

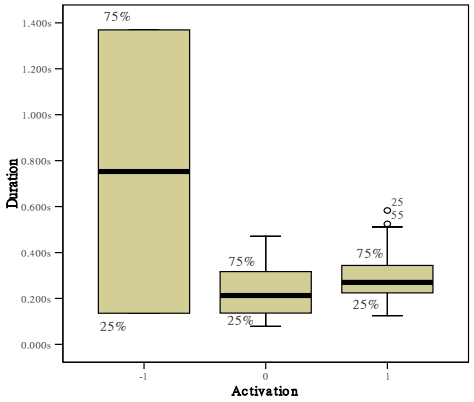


Figure 7. The duration distribution of the breath segments by radio hostess in different activations

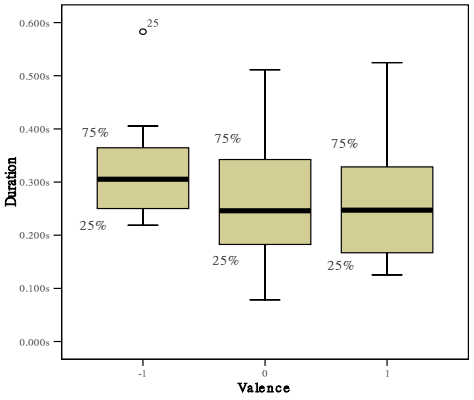


Figure 8. The duration distribution of the breath segments by radio hostess in different valences

Figures 9 and 10 indicate that the duration range of the breath segments produced by the audience is affected by the valence and activation.

From these four figures, one can get the duration of breath segments when valence and activation are 1,-1 or 0 in spontaneous speech, whose results can be used in the following perceptual experiment.

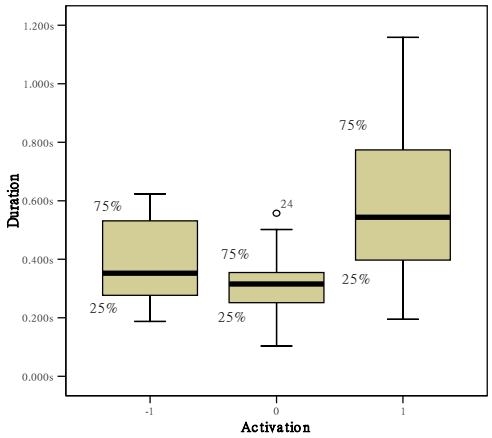


Figure 9. The duration distribution of the breath segments by audience in different activation

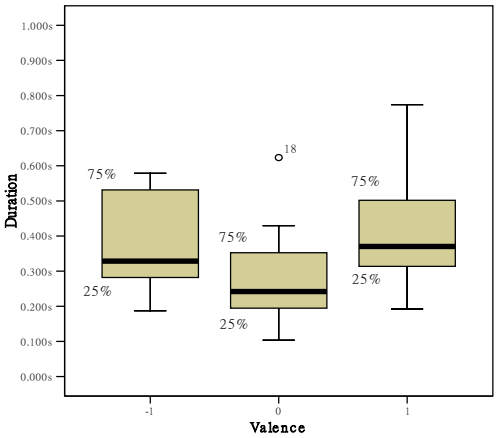


Figure 10. The duration distribution of the breath segments by audience in different valence

4.3 Rules for Inserting Breath Segments in Spontaneous Speech

The insertion rule in spontaneous speech is more complicated than that in read speech. In spontaneous speech, the breath segments will be divided into two types according to their functions: the physiological activity and the expression of emotion or attitude. The following

rules can be used in breath insertion for synthesizing spontaneous speech.

A. Physiological breath insertion without emotion is the same as that in read speech as described above. However, in dialogs there is some turn-taking. Sometimes, the breath appearing at the turn taking may overlap with the words spoken by the interlocutor or appear close to the boundary of the turn taking.

B. When the activation is -1, the duration of breath segment is set randomly from 0.2 to 0.6 second. When the activation is 1, the duration of breath segment is set randomly ranging from 0.1 to 0.4 second. When the activation is 0, the duration of breath segment is set randomly from 0.2 to 0.5 second.

C. When the valence is -1, the duration of breath segment is set randomly from 0.1 to 0.4 second. When the valence is 1, the duration of breath segment is set randomly from 0.2 to 0.5 second. When the valence is 0, the duration of breath segment is set randomly from 0.2 to 0.6 second.

D. Between every breath segment and the prosodic phrase following this breath segment there is a silence.

5. Perceptual Experiments

5.1 Stimuli

A pilot perceptual experiment is conducted to test the obtained results. The texts are selected from a read story and spontaneous dialogs. The original synthesized speech is produced by using the synthesizer provided by iFLYTEK. After that, breath segments are inserted into the synthetic speech, based on the previous rules.

Twenty subjects recruited to join the perceptual experiment are asked to judge the differences between the speech materials with and without breath for both the original and the synthesized speech. The perceptual process consists of two steps: first, the subjects are asked to compare the speech from the read story. Then, these subjects are required to perceive the breath effect in the synthesized dialogs.

Speech fragments from a read story (Little Red Hat) are numbered as X-1 (the original speech), X-2 (the original speech minus the breath segments), X-3 (the synthetic speech) and X-4 (the synthetic speech inserted with breath segments). For speech based on the spontaneous speech scripts, the two stimuli are numbered as Y-1 and Y-2, which are synthesized speech and inserted with breath segments.

5.2 Results

In the first experiment, the whole speech or segmented clips are compared. Five clips are segmented for each X. Totally, 20 clips are attained for X1, X2, X3 and X4 by segmenting at the same text boundaries. Subjects are asked to listen to and compare all counterparts with and without breath segments to judge if they are different or not and which is more natural. The subjects are only allowed to listen to the stimuli a maximum of 3 times.

The results are listed in Table 10, in which 1 stands for the counterparts (with and without breath segments) which are different, 0 means there is no difference between the perceived counterparts. 70% subjects fail to distinguish between X1 and X2. Carefully comparing X3 with X4, subjects can perceive their differences, and feel that X-4 is more natural. When smaller fragments are compared, only 38% (38 out of 100 times) can be perceived with discrepancy. The results on X3 and X4 are slightly higher, reaching 92% (92 out of 100 times). This experiment reveals that when one changes the parameters of breath segments, such as their duration, intensity and position, most of the subjects are able to perceive the differences between the original and the breath insertion speech.

Table 10. The perceptual results of the first experiment based on reading story

Subjects	X-1 and X-2 (in five clips)	X-3 and X-4 (in five clips)
1	2/5	5/5
2	5/5	5/5
3	5/5	5/5
4	2/5	5/5
5	1/5	4/5
6	1/5	5/5
7	0/5	4/5
8	1/5	4/5
9	2/5	5/5
10	2/5	4/5
11	2/5	4/5
12	1/5	5/5
13	2/5	5/5
14	3/5	4/5
15	2/5	5/5
16	2/5	4/5
17	1/5	5/5
18	1/5	4/5
19	2/5	5/5
20	1/5	5/5
Total	38/100	92/100

Table 11. The result on spontaneous dialogues Y1 and Y2

Subjects	Y-1			Y-2		
	breath	naturalness	expressiveness	breath	naturalness	expressiveness
1	1	1	0	1	1	1
2	0	0	0	1	1	0
3	1	1	0	0	0	0
4	0	0	0	1	0	0
5	0	0	0	0	0	0
6	1	0	1	1	0	0
7	0	0	0	1	1	0
8	1	0	0	0	0	0
9	0	0	0	0	0	0
10	1	1	1	1	1	1
total	5/10	3/10	2/10	6/10	4/10	2/10

The second experiment is rather simple, compared to the first one. The subjects are asked to judge which group of the two dialogs Y1 and Y2 has breath segments. If the subjects can tell the difference, they have to judge whether or not the breath segments insertion can increase the naturalness and the expressiveness. The result is shown in Table 11. The rates of breath insertion recognition are 50% and 60% for Y1 and Y2 respectively, but only 20% for expressiveness and 30% to 40% for naturalness.

6. Conclusion

This paper, with a statistical analysis made on breath segments in read and spontaneous speech, proposes some preliminary principles for inserting breath segments in synthesized speech. These principles or rules can help one better understand the physiological and expressive features in speech synthesis. Though the authors got relatively limited results in the perceptual experiments, it proves that non-verbal information is not just a simple physiological breath; instead, it is an essential element in transmitting expressiveness or attitude. In this regard, future studies should focus on other frequently encountered paralinguistic and nonlinguistic information, so that further steps may be achieved in understanding breath segments by classifying valence into more categories.

References

Alku, P., and E. Vilkman, “A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers,” *Folia phoniatrica et logopaedica* Karger ,48(55), 1996, pp. 240-254.

Cahn, J.E., "Generating Expression in Synthesized Speech," Master' s Thesis, MIT,1989.

Campbell, N., "Perception of Affect in Speech - towards an Automatic Processing of Paralinguistic Information in Spoken Conversation," In *Proceeding of 8th International Conference on Spoken Language Processing,* Jeju, Korea, 2004, pp. 881-884.

Campbell, N., "Where is the Information in Speech?" In *Proceedings of the Third ESCA/COCOSDA International Workshop*, 1998, Australia, pp. 17-20.

Chen, X.-X., A.-J. Li, et. al, "Application of SAMPA-C in SC," In *Proceeding of ICSLP2000*, 2000, Beijing, pp.VI 652-655.

Heuft, B., T. Portele, and M. Rauth, "Emotions in time domain synthesis," In *Proceeding of 4th International Conference on Spoken Language Processing*, Philadelphia, USA,1996, pp. 1974-1977.

Iida, A., N. Campbell, S. Iga, F. Higuchi, and M. Yasumura, "A Speech Synthesis System with emotion for Assisting Communication", In *Proceeding of ISCA Workshop on Speech and Emotion*, Northern Ireland , 2000, pp. 167–172.

Li, A.-J., "Chinese Prosody and Prosodic Labeling of Spontaneous Speech" In *Proceedings of International. Workshop on Speech Prosody*, Aix-en-Provence, France, 2002, pp. 39-46.

Trouvain, J, "Segmenting Phonetic Units in Laughter, Conference of the Phonetic Sciences," In 15th. *International Conference of the Phonetic Sciences (ICPhS)*, Barcelona, Spain, pp. 2793-2796.

Vroomen, J., R. Collier, and S. Mozziconacci, "Duration and intonation in emotional speech," In *Proceedings of the Third European Conference on Speech,* Berlin, Germany, 1993, pp. 577–580.

Computational Linguistics and Chinese Language Processing
Vol. 12, No. 1, March 2007, pp. 33-48
© The Association for Computational Linguistics and Chinese Language Processing

Affective Intonation-Modeling for Mandarin Based on PCA

Zhuangluan Su*, and Zengfu Wang*

Abstract

The speech fundamental frequency (henceforth F0) contour plays an important role in expressing the affective information of an utterance. The most popular F0 modeling approaches mainly use the concept of separating the F0 contour into a global trend and local variation. For Mandarin, the global trend of the F0 contour is caused by the speaker's mood and emotion. In this paper, the authors address the problem of affective intonation. For modeling affective intonation, an affective corpus has been designed and established, and all intonations are extracted with an iterative algorithm. Then, the concept of eigen-intonation is proposed based on the technique of Principal Component Analysis on the affective corpus and all the intonations are transformed to the lower-dimensional eigen sub-space spanned by eigen-intonations. A model of affective intonations is established in the sub-space. As a result, the corresponding emotion (maybe a mixed emotion) can be expressed by speech whose intonation is modified according to the above model. The experiments are performed with the affective Mandarin corpus, and the experimental results show that the intonation modeling approach proposed in this paper is efficient for both intonation representation and speech synthesis.

Keywords: Eigen-Intonation, Affective Speech, Mixed Emotion, F0 Contour, Speech Synthesis

1. Introduction

Speech can convey not only literal meanings, but also the mood and emotion of a speaker. Some researchers have proven that the contour of the speech fundamental frequency (henceforth F0 contour) plays an important role in expressing the affective information of an utterance. It is concluded that some statistical characteristics of F0 play the most important roles in emotion perception [Tao and Kang 2005]. Especially, F0 contours differ from each

* Department of Automation, University of Science and Technology of China, Hefei 230027, China
E-mail: zfwang@ustc.edu.cn
The author for correspondence is Zengfu Wang.

[Received September 16, 2006; Revised October 10, 2006; Accepted November 1, 2006]

other because of the speaker's different emotion in Mandarin [Yuan *et al.* 2002]. Due to significance of F0, the F0 contour modeling is one of the key issues that should be addressed.

The most popular F0 modeling approaches mainly use the concept of separating the F0 contour into a global trend and local variation [Abe and Sato 1992; Bellegarda *et al,* 2001]. Mandarin is a tonal language including four basic tone types and a so-called 'light' tone. The F0 contour is composed of three elements [Zhao 1980]: the tone of the syllable, the variety of tone in continuous utterance, and the movement influenced by mood. How to extract tones and intonations from speech is a difficult problem. Tian and Nurminen have proposed a data-driven tone modeling approach to describe the tonal element [Tian and Nurminen 2004]. In previous work [Su and Wang 2005], the authors of this paper also proposed an affective-tone modeling approach for Mandarin to separate F0 contour into two elements: variational tones based on syllables and intonations for prosody phrases.

In this paper, the authors propose a data-driven intonation modeling approach based on Principal Component Analysis (henceforth, PCA [Fukunaga 2000]). For modeling affective intonations, an affective corpus of Mandarin has been designed and the corresponding intonations are extracted with an iterative algorithm from the original speech. The eigen-intonation concept is proposed based on the principal components of the above intonations obtained from the affective corpus, and all the intonations are then transformed into the sub-space spanned by the eigen-intonations. The distribution of affective intonations corresponding to an emotion in the above sub-space is a help to establish the corresponding affective intonation model. As a result, speech whose intonation is modified according to the model can express the corresponding emotion, even mixed emotions. In addition, the authors will also show emotion perception results using the proposed modeling approach.

The remainder of the paper is organized as follows. The speech corpus and some statistic results of F0 based on the database are described first. Then, the algorithm of eigen-intonation extraction is described, and some of the basic properties of the eigen-intonation representation are concluded. Next, how to model the affective intonation is discussed. Last, the performance of the proposed modeling approach is given by experimental results.

2. Speech Corpus and Statistic Results of F0

Carrying on the affective speech research, a reasonable classification of the emotion is needed first, and then the speech features with different emotions can be analyzed effectively. In emotional psychology, Robert Plutchik proposed a four pair emotional ring constructed of eight pure emotions, including anger, joy, acceptance, surprise, fear, sadness, hatred and expectation. In the affective speech research for Mandarin, four emotions are generally selected, either including anger, joy, fear, sadness [Yuan *et al.* 2002; Tao and Kang 2005], or including anger, joy, surprise and sadness [Zhao *et al.* 2004]. In contrast, five emotions are selected for this paper, and

they are anger, joy, surprise, fear and sadness.

What is discussed in this paper is the global variety of the F0 contour, so a reasonable duration of the target needs to be considered. Due to the multi-level structure of prosody [Abney 1995; Li *et al.* 2000], a complicated sentence with many syllables can be divided into several simple prosody units with fewer syllables at prosody boundaries. So, studying intonation based on prosody units can transform this complicated problem into several simple ones. Moreover, it is known that prosodic phrases can keep a relatively stable intonation pattern. Therefore, the authors model intonation based on prosodic phrases in the paper.

It is known that F0 contour is influenced by several factors, including syntax, stress, speaker's emotion and his or her individual character. This paper focuses on the movement of intonation caused by emotion, and the influence of other factors such as syntax, stress, and the individual characters will not be considered. Currently, there are no effective methods that can eliminate the influence of these factors from the original speech signals directly, so the corpus used in the paper are obtained in such a way as to avoid these interferential factors' influence.

To avoid unwanted factors' influence and to simplify the following processing, the corpus is designed with some limitations. The authors have designed 40 sentences with different literal contents for the following test, and each sentence only consists of three components: subject, verb, and object. Furthermore, the subject, verb, and the object are all designed to be disyllabic words. So, each sentence only has 6 syllables in this case, and all of these sentences have the same syntax. As the length of a prosodic phrase is approximately six syllables [Zhao *et al*. 2002], each sentence consists of only one prosodic phrase. An example of such a sentence is given by "北京召开奥运". This design can be advantageous to the following experiments, and the model will be established directly based on one sentence. Each sentence is then performed by a female actor with all six emotions, including fear, sadness, neutral, anger, joy and surprise. In the end, the corpus used for analysis contains 240 total sentences, consisting of 1,440 syllables from a single speaker, with same syntax and the same individual characters. The speech signals are digitized at 16 kHz with 16-bit precision.

To evaluate the representational ability of the corpus, some experiments about the distributions of F0 are performed. Here, the F0 of a speech is extracted by using a modified autocorrelation algorithm. The results are demonstrated in Figure 1.

Figure 1 shows that "surprise", "happy" and "angry" make a very high F0, while "sad" generates lower value than the neutral state. It can also be found that the varying range of "sad" is smaller than the others. F0 parameters of "fear" make quite similar behaviors as "sad". "Angry", "happy", and "surprise" also behave similarly. All of the results accord with the conclusions given by other researches [Yuan *et al.* 2002; Zhao *et al.* 2004; Tao and Kang 2005]. So the speech corpus is representational and effective for the following analysis.

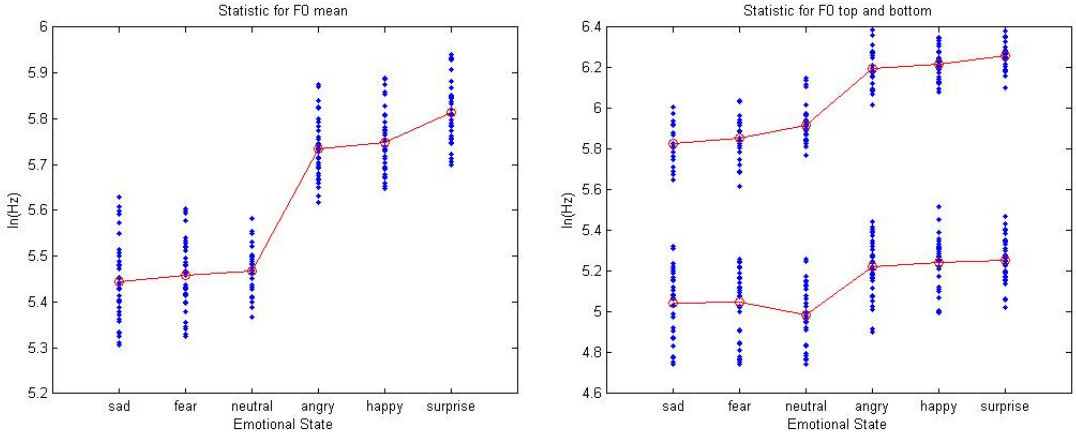


Figure 1. Statistic results for F0 with different emotional states

3. Concept of Eigen-Intonation

The affective intonation will be modeled with a concept called "eigen-intonation". The concept of eigen-intonation is derived through the use of the PCA technique. PCA [Fukunaga 2000] is a multivariate analysis method that carries out a compact description of a data set. In a PCA process, a set of correlated variables is transformed into a set of uncorrelated variables that are ordered by reducing variability, and these new uncorrelated variables are linear combinations of the original variables. It can be concluded that the first new variable contains the greatest amount of variation; the second contains the next greatest residual variance and orthogonal to the first, and so on. Thus, the last of these variables can be removed with a minimal loss of real data.

With the affective corpus in the paper, the speech intonations for sentences should be very similar in all configurations, and they should be able to be described by some "basic intonations". From the previous description, one knows that one of the main functions of PCA is that it can be used to extract new uncorrelated features from original data. According to these ideas, one can find the "basic intonations" that best account for distribution of speech intonations within the entire intonation space using the principal components analysis. The "basic intonations" are called "eigen-intonations".

With eigen-intonation, original intonations can be transformed to corresponding representations with lower dimensions. Some rules can also be possibly given out in the low-dimensional space. Moreover, the resultant rules with low dimensions have simpler expression, and it is advantageous to control the rules for the goal of this study.

4. Analysis for Eigen-Intonation

The concept of eigen-intonation is proposed based on PCA technique. Mathematically, the principal component analysis involves an eigen analysis on a covariance matrix. A good low-dimensional representation in the space of possible speech intonations can be achieved by considering only a few principal components or eigenvectors, corresponding to the first largest eigenvalues.

4.1 Extraction of Intonation

In order to obtain the intonation of a speech, the F0 contour of the speech should be extracted first. After that, the F0 contour will be separated into a global variety, which is regarded as intonation, and rapidly-varying components corresponding to local changes based on syllables. The details of intonation extraction are described in the following.

The entire intonation extracting algorithm can be divided into five main steps:

1) Estimating initial F0 values based the modified normalized autocorrelation from voiced regions of the original speech.
2) Cubic Hermite interpolating for unvoiced regions and obtaining a continuous F0 curve.
3) Filtering the continuous F0 contour with two serial modified smoothing processes.
4) Applying piecewise three-order polynomial iterative fitting to the entire F0 contour, the n-th iterative processing step is as:
   (a) Fitting the entire F0 contour with n pieces of cubic polynomial.
   (b) Calculating the fitting error E_n.
   (c) If $E_n < E_t$, ending the iterative algorithm and taking n pieces of cubic polynomial fitting as final resultant F0 contour. Else, $n = n + 1$, go to (a). Where E_t is a given threshold of maximal fitting error.
5) The ln(F0) contour is passed through a high-pass filter with a stop frequency at 0.5Hz, and the residual low frequency contour after filtering is denoted as L_F contour.

From the authors' previous work [Su and Wang 2005], The L_F contour can be regarded as the F0 global variety of a speech. As all sentences have the same syntax and each sentence consists of only one prosodic phrase in this corpus, the model can be established directly based on one sentence. It is to say that the resultant L_F contour of the algorithm for each sentence in the corpus is the modeling target, intonation based prosodic phrase (henceforth intonation). Finally, each intonation is normalized into an N-dimensional vector (N = 100 in the paper).

4.2 PCA for Intonation

Let the data set of intonations be $I_1, I_2, \ldots I_M$, where I_i is an N-dimensional intonation sample, and M is the number of intonations (M = 240 in the paper). Then the intonation covariance matrix $C_{N\times N}$ is computed by (1).

$$C = \frac{1}{M}\sum_{i=1}^{M}(I_i - m)(I_i - m)^T \tag{1}$$

Where, m is the average intonation calculated by (2).

$$m = \frac{1}{M}\sum_{i=1}^{M} I_i \tag{2}$$

The differential intonations matrix A is defined as (3).

$$A = \frac{1}{\sqrt{M}}\left[I_1 - m, I_2 - m, \ldots, I_M - m\right] \tag{3}$$

Then, $C = AA^T$ is an $N\times N$ covariance matrix. The eigen analysis on the covariance matrix $C_{N\times N}$ yields a set of positive eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ in descending order and the corresponding eigenvectors, $\{V_1, V_2, \ldots, V_N\}$. The first L ($L < N$) eigenvectors, denoted as $U = \{V_i, i = 1, 2,\ldots, L\}$, are selected as principal components, and the intonations corresponding to these L vectors are so-called eigen-intonations, denoted as U_o.

The eigen sub-space spanned by the principal components U is called sub-space of intonation, denoted as P, and the original space of intonation is denoted as O. All intonations in O can be projected to be the corresponding representations in P. It is known that the dimension of P is lower than that of O, and one can establish the rules of intonation in P and then restore the resultant intonations in O. Obviously, rules with lower dimension are easily controlled. Next, restoration of intonation will be discussed.

4.3 Restoration Based on PCA

According to the principal component analysis, the original intonations in O are projected into the sub-space P as (4).

$$\Omega_k = U^T(I_k - m), \quad k = 1, 2, \ldots, M \tag{4}$$

Where, Ω_k is coordinate vector of the k-th intonation. With Ω, the intonation samples are restored as (5), and the final approximation of the original intonations I is given out as (6), denoted as J.

$$B = U\,\Omega \tag{5}$$

$$J_k = B_k + m \quad k = 1, 2, \ldots, M \tag{6}$$

Especially, let $B = U$ in (6), intonations corresponding to U can be given out, and that are eigen-intonations U_o. It can be concluded that although U_o is higher than U, the configuration of U_o is same as U. So the authors do not distinguish them when their configurations are discussed.

To evaluate the ability of restoration, the restoring rate for k-th intonation is defined as (7).

$$R_k = 1 - \frac{\|I_k - J_k\|}{\|I_k\|}, \quad k = 1, 2, ..., M \tag{7}$$

The final restoring rate of the entire algorithm is defined as (8).

$$r = \left\{ \sum_{k=1}^{M} R_k / M \right\} \times 100\% \tag{8}$$

5. Affective Intonation

5.1 About Affective Intonation

Affective intonation is the concept that a speech with a certain affective intonation can express a corresponding emotion. Some works of speech prosody have proposed much qualitative analysis for affective intonations, and this paper will try to give quantitative affective intonation rules. At last, speech whose intonation is modified according to a certain affective intonation obtained in the paper can express the corresponding emotion.

In order to research affective problems, emotion can be classified. Robert Plutchik [Plutchik 1960] considered that the emotions felt in normal human life were complicated and mixed, and considered some intensity of the eight pure emotions constructing a mixed emotion. So, in a similar way to him, all the mixed-emotional intonations are supposed to be defined by some vectors in the form of linear combination of the coefficients in the paper, where the vectors are the principal components U and the coefficient is the coordinate vector Ω_k in (4). Based on this assumption, one can easily change the coefficient corresponding to a certain eigen-intonation to control some configuration of final affective intonation for the goal. How to perform the assumption is discussed in the following.

5.2 Modeling Affective Intonation

Let the set of emotions be a, $a = 1, 2,\ldots,6$ representing anger, joy, surprise, fear, sadness and neutral emotional state. Intonations extracted from the speeches with emotion a are denoted as N-dimensional vector I^a in original space O. Let $I = I^a$ in (4), and I^a be projected into the sub-space P, denoted as Ω^a. Ω^a is distributed in different regions in P for the different emotions a, and the mass kernel vectors $\bar{\Omega}^{\alpha}$ are computed as (9).

$$\overline{\Omega}^{\alpha} = \sum_{k=1}^{K_\alpha} \Omega_k^{\alpha} / K_\alpha, \quad \alpha = 1, 2, ..., 6 \tag{9}$$

Where, Ω_k^{α} is the projecting representation vector (henceforth projecting vector) in P of the k-th intonation with emotion a. K_a is the total number of all intonation samples with emotion a.

$\{\overline{\Omega}^{\alpha}, a = 1, 2, \ldots, 6\}$ are the resultant affective intonations with low dimension basing eigen-intonation. They are restored in the original intonation space O as (10).

$$T_\alpha = U\overline{\Omega}^{\alpha} + m, \quad \alpha = 1, 2, ..., 6 \tag{10}$$

Where T_a are the final affective rule-intonations (henceforth rule-intonations) and they can be applied directly to modify the target intonation for synthesizing affective speech, which will be performed in the following experiments.

6. Experimental Results and Discussion

6.1 Analysis on Eigen-Intonation

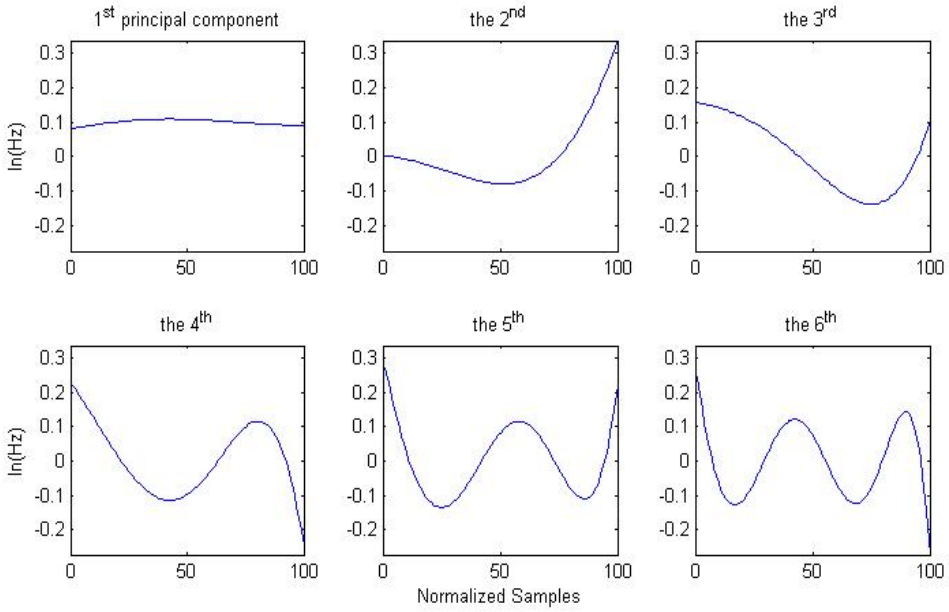


Figure 2. Eigen-intonation of the affective speech

To demonstrate the eigen-intonations, a PCA experiment using the affective speech corpus was performed. The first six principal components U are shown in Figure 2 and the authors do not distinguish the principal components selected and eigen-intonations here. It can be seen that the varying range of the first component is the smallest, and it is also the highest. So the

first eigen-intonation represents the flat and positive pitch. The second eigen-intonation contributes a big rising component, and the third matches a falling intonation with a little rising at the end. The fourth can be viewed as adding a falling part to the end of the third. The varying ranges are same between the fifth and the sixth, and their global trends are flat with big rising and falling varying. These two can be viewed as adding a rising or a falling part to the end of the previous component. It will be known that the sixth component contains a very small contribution of energy or variance to the intonation contour in the following analysis.

Based on the previous resultant eigen-intonations, the authors carry out the restoring experiment using L components selected, respectively considering L be 3, 4, 5 and 6. The results are shown in Table 1.

Table 1. The restoring rate r with L components selected

L – component number	3	4	5	6
r – restoring rate	81.61%	95.71%	99.46%	99.89%

From Table 1, it can be concluded that selecting five components is acceptable, but with six principal components, the restoring rate is 99.89% and the approximation error is almost equal to zero. The approximating examples are shown in Figure 3. That means a good six-dimensional representation for the space of all speech intonations is achieved, and these eigen-intonations are very efficient for intonation representation.

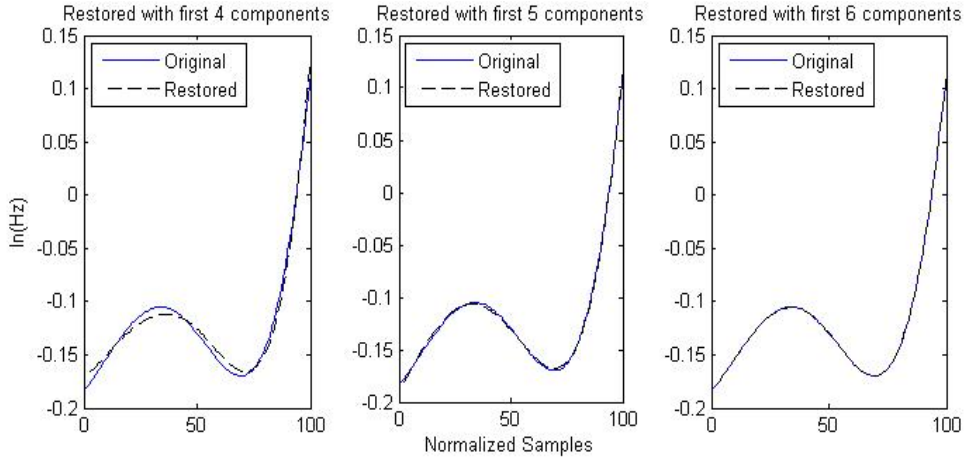


Figure 3. Illustration for restoring with eigen-intonations

6.2 Modeling Affective Intonation

The emotional state expressed by intonation of each affective speech in the corpus is known, and there are six categories of emotions, including the neutral state. And there are 40 speeches within each emotional state. According to Section 5.2, all affective intonations labeled with

different emotions are projected into six-dimensional sub-space P spanned by eigen-intonations. The distribution of first three weights of the projecting vector Ω^a is shown as Figure 4, and the mass kernel of each emotional state is indicated by red color in the figure.

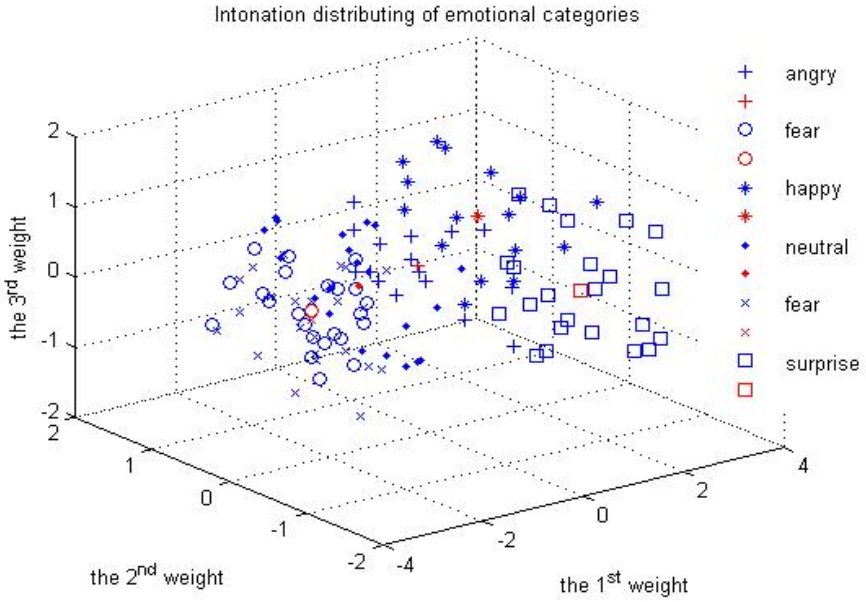


Figure 4. Distribution of first 3 weights of affective intonations in eigen sub-space

From Figure 4, one can see that the kernel of surprise, job and anger is far from that of neutral, where the "surprise" is farthest and then "angry" is next. However, the "fear" almost distributes in the same region with "sad", and they can be distinguished from the neutral emotional region. In addition, it can be known from analysis on eigen-intonations that the last several weights corresponding to these three weights in the figure contain a very small contribution of energy or variance, so the difference of their distribution is not as clear as in Figure 4.

Now the projecting vectors Ω^a in P of original intonations labeled with emotion are given out as well as the corresponding kernel vector $\bar{\Omega}^{\alpha}$ for each emotional state. By restoring with eigen-intonations, the kernel vectors are transformed as (10) into the original space, there they are regarded as rule-intonations. The rule-intonations representing emotion states are illustrated in Figure 5. From the figure, one can see that the intonations of anger, job and surprise are high, where the variety of surprise is greatest. However, the "fear" is flat and low, similar to that of the "sad". All these qualitative results are in line with the previous works of other researchers. So the resultant rule-intonations are efficient for expressing emotions in theory.

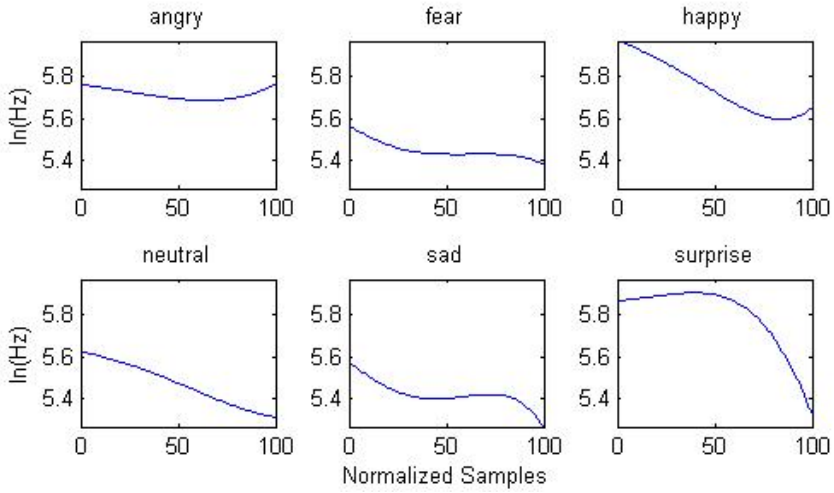


Figure 5. Affective rule-intonations T_α

6.3 The Mixed-Emotional Intonation

When the affective rule-intonation was modeled with eigen-intonation in the previous sub-section, the emotion labeled in the corpus and expressed by resultant intonation was supposed to be pure. It is known that the emotions of humans felt in normal life are not always so simple, and they are usually mixed with several so-called pure emotions, whose intensities differ corresponding to constructing the different emotions. The experiment is performed as the following to explain that the modeling approach proposed with eigen-intonation is also effective for representing the mixed-emotional intonation.

All affective intonations labeled with emotions have been projected into sub-space P and the distribution of first three weights of the projecting vector Ω^a in P has been shown in Figure 4. Now only the mass kernel of each emotional state, which is corresponding to the resultant rule-intonation, is represented in Figure 6.

Nine equal space points in line between the neutral kernel and the surprise kernel are selected and indicated in the figure. If the kernel explains pure emotions, then what the points selected explain are the mixed emotions. Along the arrow in Figure 6, points at the starting vertex explain more neutral and those at the ending vertex explain more surprise. So the emotions expressed by the intonations correspond to these points transfer from neutral to surprise along the arrow and they are mixed. The mixed-emotional intonations corresponding to the selected-points are restored in original space and shown in the left of Figure 7. It can be concluded from the figure that, along the arrow, the first rule-intonations can express more neutral and the last ones express more surprise and all of them express the mixed emotions.

Another nine equal space points between the happy kernel and the surprise kernel are also selected and the same experiment is performed. The illustrations of the experiment are shown in Figure 6 and the right of Figure 7.

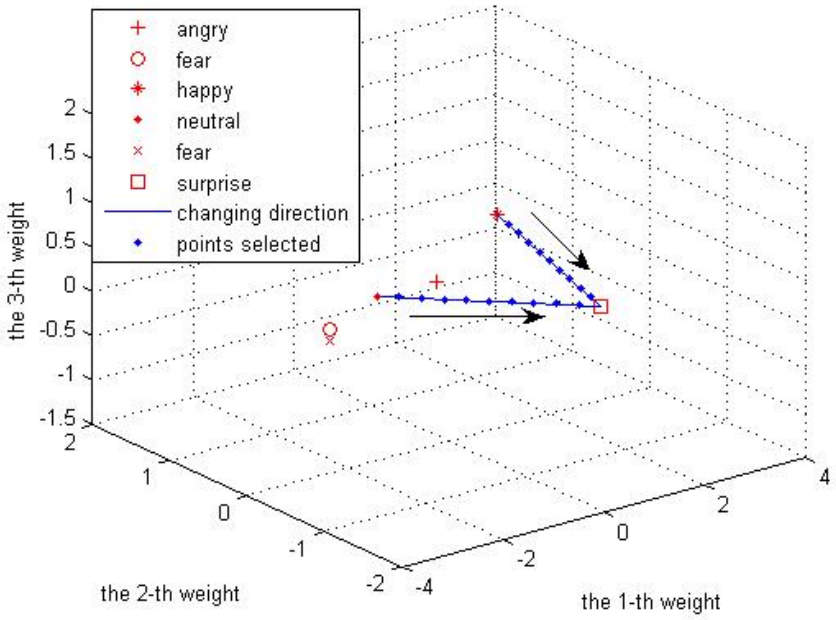
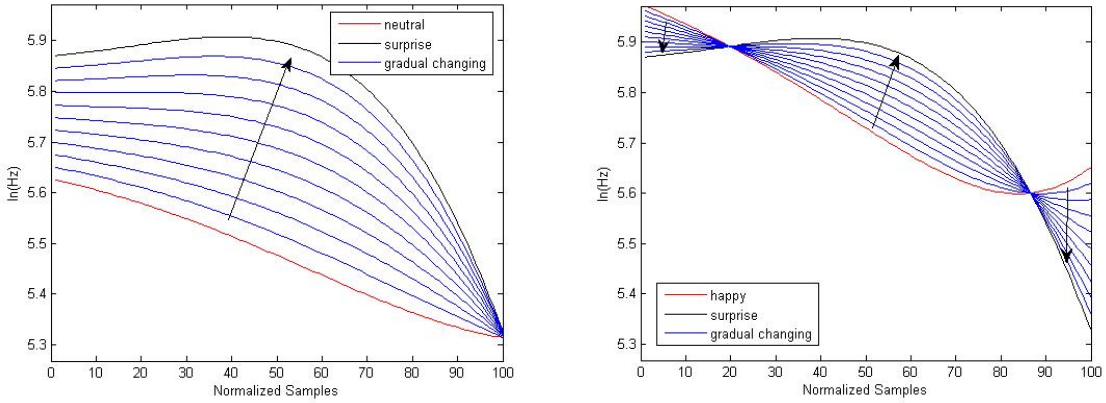


Figure 6. Transferring illustration of affective intonation in sub-space



Note:

The arrows in the figure indicate the gradual varying direction corresponding to that in sub-space showed in Figure 6 and each gradual changing curve is corresponding to one point selected in Figure 6.

Figure 7. Intonations transferring corresponding to that in sub-space

Figure 6 and Figure 7 show that the mixed-emotional intonation can be represented with eigen-intonation, so one can control the relative position of intonation-representation in the sub-space to explain the certain mixed-emotion felt in the usual human life. To sum up, the modeling approach proposed with eigen-intonation is effective for representing not only the simple emotional intonation but also the mixed-emotional intonation.

6.4 Synthesis with Affective Intonation

Based on the linear predictive coding technology [Quatieri 2004], the authors analyzed neutral speeches, modified their intonations with the six rule-intonations, respectively, and re-synthesized them. For example, the intonation of a neutral speech is modified to the surprise intonation, and the demonstration is shown as Figure 8. In the figure, the top is the waveform of the neutral speech, and the bottom includes the original F0 contour, the original intonation, the modified intonation, and the resultant F0 contour of the neutral speech. Moreover, the intonation of an original surprise speech is also plotted in the bottom figure for contrast. Figure 8 shows that the modified intonation is similar to the original intonation of the surprise speech, and the resultant F0 contour is higher than expressing surprise.

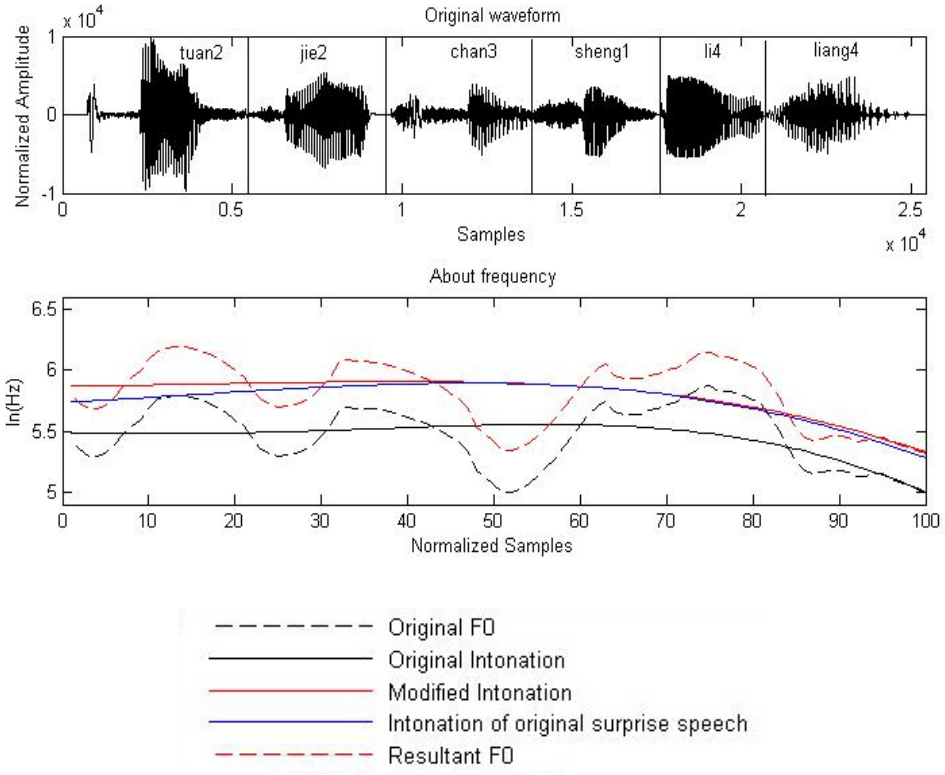


Figure 8. Illustration for modifying intonation with surprise rule-intonation

In the perception experiment, the listener was asked to judge the emotional state of the speech sound. The results show that, though it is difficult to distinguish anger from happy, and also can not point out whether the speech sounded closer to fear or sadness, it is easy to tell the emotional states such as joy, surprise, and fear of one speech. So one can conclude that the rule-intonations are almost corresponding to the emotional state and the eigen-intonation modeling method is efficient.

7. Conclusion

The F0 contour plays an important role in expressing the affective information of an utterance, and the most popular F0 modeling approaches are mainly using the concept of separating the F0 contour into a global trend and local variation. Mandarin is a tonal language, and the global trend of F0 contour is caused by speaker's mood and emotion, which is focused on in this paper, and that is called affective intonation. Affective intonation is the concept that a speech with a certain affective intonation can express a corresponding emotion. Some works of speech prosody have proposed much qualitative analysis for affective intonations, and the paper has given out quantitative rule-intonation.

In order to establish the model of affective intonation, an affective corpus of Mandarin was obtained with some limitation for affective research goal and all intonations were extracted from the original speeches. Then the eigen-intonation concept was proposed basing PCA on the affective corpus and all the intonations were transformed to lower-dimensional representations in the eigen sub-space spanned by eigen-intonations. A model of affective intonations was established in the sub-space and then was restored in the original space of intonation to form the rule-intonations. As a result, speech whose intonation is modified according to a certain rule-intonation can express the corresponding emotion, even the mixed emotion.

The authors have performed experiments with the affective Mandarin corpus. And the experimental results are in line with the theoretical analysis and the intonation modeling approach proposed is proved to be efficient for representing the simple emotional and mixed-emotional intonation. Future work will focus on how to accurately give out the boundaries of the pure emotional regions in sub-space with eigen-intonation.

Acknowledgements

This work is supported by Open Foundation of National Laboratory of Pattern Recognition, China. The authors would like to thank Dr. Tieniu Tan and Dr. Jianhua Tao for their help.

References

Abe, M., and H. Sato, "Two-stage F0 control model using syllable based F0 units," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, San Francisco, USA, 1992，pp.53-56.

Abney, S., "Chunks and dependencies: Bringing processing evidence to bear on syntax," in *Jennifer Cole and Georgia Green and Jerry Morgan(Eds.): Computational Linguistics and the Foundations of Linguistic Theory*, pp. 145-164, CSLI, 1995.

Bellegarda, J., K. Silverman, K. Lenzo, and V. Anderson, "Statistical prosodic modeling: from corpus design to parameter estimation," *IEEE Trans. Speech and Audio Processing*, 9(1), 2001, pp. 52-66.

Fukunaga, K., *Introduction to statistical pattern recognition*, Academic Press, Dordrecht, 2000.

Li, A., M. Lin, X. Chen, Y. Zu, G. Sun, W. Hua, Z. Yin, and J. Yan, "Speech corpus of Chinese discourse and the phonetic research," in *Proceedings of Sixth International Conference on Spoken Language Processing, 2000, Beijing, China,* pp. 13-18.

Plutchik, R. "The multifactor-analytic theory of emotion," *Journal of Psychology*, 50, 1960, pp. 153-171.

Quatieri, T. F., *Discrete-Time Speech Signal Processing: Principles and Practice*, House of Electronics Industry, Beijing, 2004.

Su, Z., and Z. Wang, "An Approach to Affective-Tone Modeling for Mandarin," *Lecture Notes in Computer Science 3784*，ed. By J. Tao, T. Tan, and R.W. Picard, Springer, 2005, pp. 390-396.

Tao, J., and Y. Kang, "Features Importance Analysis for Emotional Speech Classification," *Lecture Notes in Computer Science 3784*，ed. By J. Tao, T. Tan, and R.W. Picard, Springer, 2005, pp. 449-457.

Tian, J., and J. Nurminen, "On analysis of eigenpitch in Mandarin Chinese," in *Proceedings of 2004 International Symposium on Chinese Spoken Language Processing*, 2004, Beijing, China, pp. 89-92.

Yuan, J., L. Shen, and F. Chen, "The acoustic realization of anger, fear, joy and sadness in Chinese," in *Proceedings of seventh International Conference on Spoken Language Processing,* 2002, Denver, Colorado, USA, pp. 2025–2028.

Zhao, L., C. Jiang, C. Zou, and Z. Wu, "A study on Emotional Feature Analysis and Recognition in Speech," *Acta Electronica Sinica*, 32(4), 2004, pp. 606-609.

Zhao, S., J. Tao, and H. Cai, "Rule-learning Based Prosodic Structure Prediction," *Journal of Chinese Information Processing*, 16(5) , 2002, pp. 30-37.

Zhao, Y., *Problems of Language*, Commercial Press of China, Beijing, 1980.

Computational Linguistics and Chinese Language Processing
Vol. 12, No. 1, March 2007, pp. 49-64
© The Association for Computational Linguistics and Chinese Language Processing

Manifolds Based Emotion Recognition in Speech

Mingyu You*, Chun Chen*, Jiajun Bu*, Jia Liu*, and Jianhua Tao+

Abstract

The paper presents an emotional speech recognition system with the analysis of manifolds of speech. Working with large volumes of high-dimensional acoustic features, the researchers confront the problem of dimensionality reduction. Unlike classical techniques, such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), a new approach, named Enhanced Lipschitz Embedding (ELE) is proposed in the paper to discover the nonlinear degrees of freedom that underlie the emotional speech corpus. ELE adopts geodesic distance to preserve the intrinsic geometry at all scales of speech corpus. Based on geodesic distance estimation, ELE embeds the 64-dimensional acoustic features into a six-dimensional space in which speech data with the same emotional state are generally clustered around one plane and the data distribution feature is beneficial to emotion classification. The compressed testing data is classified into six emotional states (neutral, anger, fear, happiness, sadness and surprise) by a trained linear Support Vector Machine (SVM) system. Considering the perception constancy of humans, ELE is also investigated in terms of its ability to detect the intrinsic geometry of emotional speech corrupted by noise. The performance of the new approach is compared with the methods of feature selection by Sequential Forward Selection (SFS), PCA, LDA, Isomap and Locally Linear Embedding (LLE). Experimental results demonstrate that, compared with other methods, the proposed system gives 9%-26% relative improvement in speaker-independent emotion recognition and 5%-20% improvement in speaker-dependent recognition. Meanwhile, the proposed system shows robustness and an improvement of approximately 10% in emotion recognition accuracy when speech is corrupted by increasing noise.

* College of Computer Science, YuQuan Campus, ZheJiang University, Hangzhou 310027, CHINA
E-mail: {roseyoumy, chenc, bjj, liujia}@zju.edu.cn
The author for correspondence is Jiajun Bu.
+ National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing 100080, CHINA
E-mail: jhtao@nlpr.ia.ac.cn

[Received August 10, 2006; Revised September 11, 2006; Accepted September 30, 2006]

Keywords: Enhanced Lipschitz Embedding (ELE), Dimensionality Reduction, Emotional Speech Analysis, Emotion Recognition

1. Introduction

Human-machine interaction technology has been investigated for several decades. Recent research has put more emphasis on enabling computers to recognize human emotions. As the most effective method in human-human and human-machine communication, speech conveys vast emotional information. Accurate emotion recognition from speech signals will benefit the human-machine interaction and will be applied to areas of entertainment, learning, social development, preventive medicine, consumer relations, etc. [Picard 1997].

The general process of emotion recognition from speech signals can be formulated as below: extracting acoustic features such as Mel-Frequency Cepstral Coefficient (MFCC), Linear Predictive Cepstral Coefficient (LPCC) or low-level features [Ververidis *et al.* 2004], reducing feature dimensionality to an appropriate range for less computational complexity and recognizing emotions with trained SVM, Hidden Markov Model (HMM), Neural Network (NN) or other classifiers.

Dimensionality reduction methods can be grouped into two categories: Feature Selection (FS) and Feature Extraction (FE). An FS method chooses a subset from the original features, preserving most characteristics of the raw data. Ververidis [Ververidis *et al*. 2004] used the Sequential Forward Selection (SFS) method to select the five best features for the classification of five emotional states. However, feature selection needs complex computation to evaluate all the features. How to acquire the best feature set is another tough task. An FE method projects the original features to a completely new space with lower dimensionality through linear or nonlinear affine transformation. PCA, LDA and Multidimensional Scaling (MDS) are popular feature extraction techniques. PCA finds a set of the most representative projection vectors such that the projected samples retain the most information about the original samples. Lee [Lee *et al*. 2002] used PCA to analyze the feature set in classifying two emotions in spoken dialogs. Chuang [Chuang *et al*. 2004] adopted PCA to select 14 principle components from 33 acoustic features in the analysis of emotional speech. LDA uses the class information and finds a set of vectors that maximize the between-class scatter while minimizing the within-class scatter. MDS computes the low dimensional representation of a high dimensional data set that most faithfully preserves the inner products between different input patterns. LDA and MDS have also been employed to reduce the feature dimensionality for emotion recognition [Go *et al*. 2003]. Though widely used for their simplicity, PCA, LDA and MDS are limited by their underlying assumption that data lies in a linear subspace. For nonlinear structures, these methods fail to detect the true freedom degrees of the data.

Recently, a number of research efforts have shown that the speech points possibly reside on a nonlinear submanifold [Jain *et al*. 2004; Togneri *et al*. 1992]. The classical ways of projecting speech into low dimensional space by linear methods are not suitable. Some nonlinear techniques have been proposed to discover the nonlinear structure of the manifold, *e.g*. Isomap [Tenenbaum *et al*. 2000] and LLE [Roweis *et al*. 2000]. Isomap is based on computing the low dimensional representation of a high dimensional data set that most faithfully preserves the pairwise distances between input patterns as measured along the submanifold from which they are sampled. The LLE method captures the local geometry of complex embedding manifold by a set of linear coefficients that best approximate each data point from its neighbors in the input space. These nonlinear methods do yield impressive results in some statistical pattern recognition applications [Jain *et al*. 2004]. However, they yield maps that are defined only on the training data points, so how to evaluate the maps on novel testing data points remains unclear. Lipschitz embedding [Bourgain 1985; Johnson *et al*. 1984] is another nonlinear dimensionality reduction method which works well when there are multiple clusters in the input data [Chang *et al*. 2004]. It is suitable for emotion classification whose input data can be grouped into several emotions.

Most previous work on detecting emotional states investigated speech data recorded in a quiet environment [Song *et al*. 2004; Zeng *et al*. 2005], but humans are able to perceive emotions even in a noisy background. The nonlinear manifold learning algorithms mentioned above [Tenenbaum *et al.* 2000; Roweis *et al.* 2000; Bourgain 1985] try to discover the underlying reason of how humans perceive constancy even though the raw sensory inputs are in flux. Facial images with different poses and lighting directions were also observed to make a smooth manifold [Tenenbaum *et al.* 2000]. Similarly, speech with different emotions, even corrupted by noise, could also be embedded into a low dimensional nonlinear manifold, although none of the previous work has paid attention to this area.

In this paper, an enhanced Lipschitz embedding system is developed to analyze the intrinsic manifold of both emotional speech recorded in quiet environment and those corrupted by noise. Geodesic distance is expected to reflect the true geometry of the emotional speech manifold. With geodesic distance estimation, ELE is developed to embed the extracted acoustic features into a low dimensional space. Then, a linear SVM is trained to recognize the emotional states of the embedded results. In addition, other dimensionality reduction methods such as PCA, LDA, feature selection by SFS with SVM, Isomap, and LLE are implemented for comparison.

The rest of the paper is organized as follows. Section 2 gives a brief description of the emotional speech recognition system. Section 3 presents the ELE algorithm. Experimental results are provided and discussed in Section 4. Section 5 concludes the paper and discusses future work.

2. System Overview

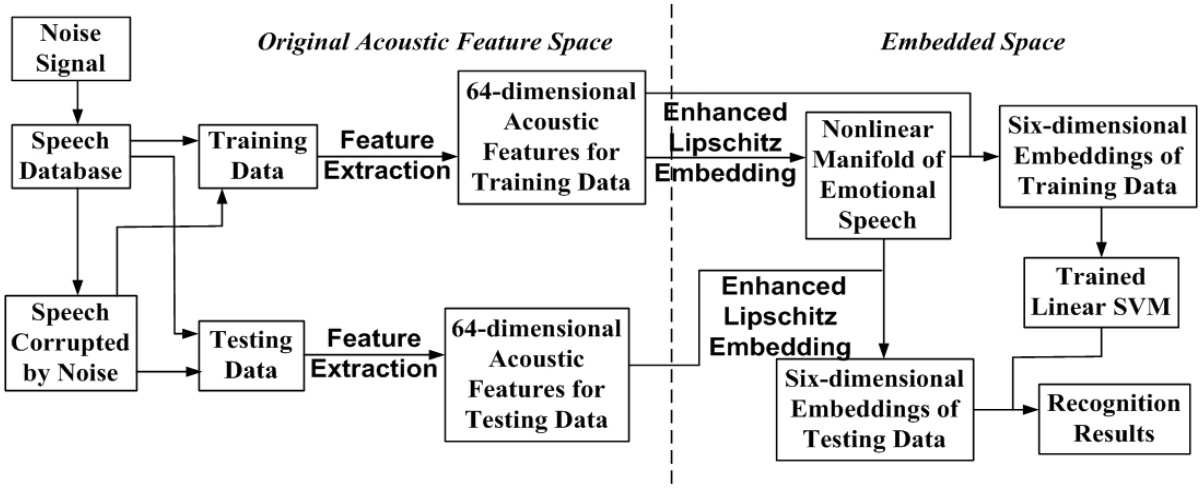


Figure 1. The Framework of Emotion Recognition from Speech

Figure 1 displays the overall structure of this system. Clean speech from the database and speech corrupted by generated noise are both investigated in the system. The emotional speech analysis is done in two phases in this system: training and testing. In the training phase, 64-dimensional acoustic features for each training utterance are obtained after feature extraction. Using ELE, a six-dimensional submanifold is then gained to embody the intrinsic geometry of the emotional training data. Finally, a linear SVM is trained by the embedded training data. In the testing phase, the feature extraction method also extracts 64-dimensional acoustic features for the testing data. The high-dimensional features are then projected into the six-dimensional manifold obtained in the training phase. The emotional state of the testing data is recognized by the trained SVM system. There are two feature spaces mentioned in the workflow: the original acoustic feature space, which is a high-dimensional space found before feature embedding and the embedded space, which is a low-dimensional space found after feature projection.

3. Enhanced Lipschitz Embedding (ELE)

A Lipschitz embedding is defined in terms of a set $R(R=\{A_1, A_2, \cdots, A_k\})$, where $A_i \subset S$ and $\bigcup_{i=1}^{k} A_i = S$. The subset A_i is termed the reference set of the embedding. Let $d(o, A)$ be an extension of the distance function d from object o to a subset $A \subset S$, such that $d(o, A) = \min_{x \in A} d(o, x)$. An embedding with respect to R is defined as a mapping F such that $F(o) = (d(o, A_1), d(o, A_2), \cdots, d(o, A_k))$. In other words, Lipschitz embedding defines a coordinate space where each axis corresponds to a subset $A_i \subset S$ and the coordinate values of object o are the distances from o to the closest element in each A_i.

The distance function d in Lipschitz embedding reflects the essential structure of data set. Due to the nonlinear geometry of the speech manifold, Euclidean distance fails to find the real freedom degrees of the manifold. Tenenbaum *et al.* [Tenenbaum *et al.* 2000] tried to preserve the intrinsic geometry of the data by capturing the geodesic distances between all pairs of data points, which is followed by the algorithm found in this research.

In this new approach, the speech corpus is divided into six subsets $\{A_1, A_2, \cdots, A_6\}$ according to six emotional states (neutral, anger, fear, happiness, sadness and surprise). Object o of speech corpus is embedded into a six-dimensional space where the coordinate values of o are obtained from the process below.

(1) Construct a graph G connecting neighbor data points. The edge length is determined by the Euclidean distance between neighbor points. The detailed operation can be formulated as Equation (1).

Initiate element m_{ij} in matrix M:

$$m_{ij} = \begin{cases} \sqrt{\sum_{\partial=1}^{64}(x_\partial - y_\partial)^2} : \forall i, j \in KNN \\ C : else \end{cases} \tag{1}$$

Here, m_{ij} stands for the geodesic distance from point i to j. $i, j \in KNN$ means that j is among the k nearest neighbors of i. Specifically, k is set to 10 in this method, which will be discussed further in the following section. i and j are data points in the 64-dimensional feature space, $i = [x_1, x_2, \cdots, x_{64}]$ and $j = [y_1, y_2, \cdots, y_{64}]$. C is a very large constant which guarantees that i and j are unconnected in the graph G consisting of speech data points. Matrix M actually corresponds to the neighborhood graph G whose edge only connects neighbor data points.

(2) Reconstruct matrix M. Replace element m_{ij} with the length of the shortest path between data point i and j in graph G. The shortest path between i and j can be found by tracing through the edges in graph G.

$$m_{ij} = \min\{m_{ij}, m_{ik} + m_{kj}\} \tag{2}$$

Matrix M contains the shortest path distances between all pairs of points in graph G constructed in Equation (1).

(3) Get the coordinate values of $o(\{o_1, o_2, \cdots, o_6\})$ from matrix M. The coordinate value of object o to axis A_i is the distance from o to the closest element in A_i.

$$o_r = \min_{\mu \in A_r} m_{o\mu} \tag{3}$$

where $m_{o\mu}$ is an element of matrix *M*. In this work, object o is projected into a space with six axes $\{A_1, A_2, \cdots, A_6\}$ in accordance with the six emotional states.

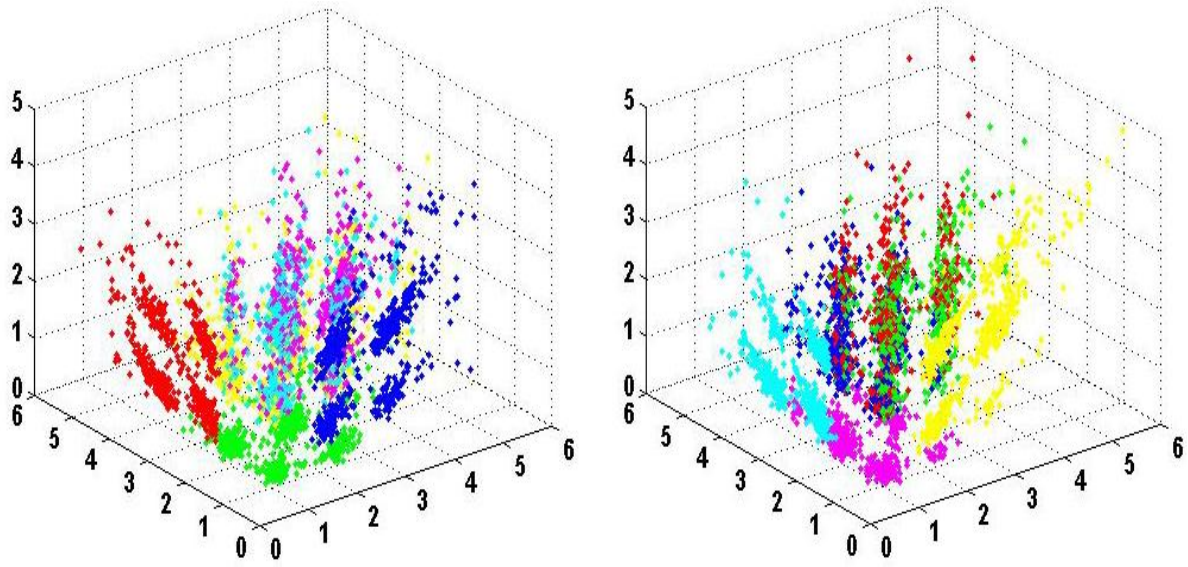


Figure 2. Training data in the embedded space. Different colors correspond to different emotions.

Figure 2 shows the six-dimensional embeddings of 64-dimensional training speech corpus in the six emotional states. Figure 2(a) reveals the first three dimensions of the embedded space and (b) displays the other three dimensions. Emotions neutral, anger and fear, denoted by points in red, green and blue, are easy to be separated in the first three dimensions. Happiness, sadness and surprise, denoted by light blue, yellow and pink are separable in the last three dimensions, though they are mixed in Figure 2(a). Actually, points of the same emotional state are highly clustered around one plane in the embedded space. The distribution property of data points in the six-dimensional space indicates that they can be easily classified into six clusters.

In the proposed ELE technique, the distance matrix *M* is constructed on training data. The training data projection easily depends on the minimal distance to each emotional speech class. Similar to Isomap and LLE, how to evaluate new testing data is still unclear. It is impossible to reconstruct matrix *M* combining the testing data because it is time consuming. Based on the constructed matrix *M*, the authors propose an approach to compute the coordinate values of testing data *t* in the embedded space.

(1) Based on Euclidean distance, the *k* nearest neighbors ($\{n_1, n_2, \cdots, n_k\}$), with distances $\{d_1, d_2, \cdots, d_k\}$, of testing data *t* are found in the training data set.

(2) Get the coordinate values $(\{v_n^1, v_n^2, \cdots, v_n^6\}_{n=1}^k)$ of the k neighbors from matrix M. The k nearest neighbors come from the training data, so their coordinates can be found with the processes mentioned in the training phase.

(3) Compute the coordinate values of testing data t $(\{t_1, t_2, \cdots, t_6\})$. In this approach, the testing data t makes the shortest paths to subsets through its neighbors. Therefore, the geodesic distances of t to subsets can be approximated by averaging the sum of "short hops" to its neighboring points and the geodesic distances of its neighbors.

$$t_i = \frac{1}{k} * \sum_{\partial=1}^{k} (d_\partial + v_\partial^i) \tag{4}$$

where k is set to 10 in the proposed system. Instead of using the minimum value, average approximation defined in Equation (4) is adopted to be the distance measurement of t for a robust performance.

4. Experiments

4.1 Speech Corpus

The speech database used in the experiment is from National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. The corpus is collected from four Chinese native speakers including two men and two women. Everyone reads 300 sentences in six emotions involving neutral, angry, fear, happy, sad and surprise. The total amount of sentences is thus 300*6*4 = 7200. The speech corpus is sampled at 16 kHz frequency and 16-bit resolution with monophonic Windows PCM format.

The clean speech data were also suppressed by generated noise signal. Gaussian white noise and sinusoid noise generated by LabVIEW were both added to the speech database at various signal-to-noise ratios (SNR) as determined by Equation (5). Gaussian white noise and sinusoid noise appear frequently in both real and research environments.

$$\eta = 10 \lg \frac{\frac{1}{n}(\sum_{i=1}^{n} x_i)^2}{\frac{1}{n}(\sum_{j=1}^{n} y_i)^2} \tag{5}$$

Where x_i is a sample from the speech signal and y_i is a sample from the noise. Due to the variations of speech signals' energy in different emotions, average SNR was measured among an individual's utterances in all emotions. The SNRs of tested noisy speech were approximately 21dB, 18dB, 15dB, 11dB, and 7dB. Noisy speech with lower SNR was excluded, due to difficulty in extracting pitch from them.

4.2 Acoustic Features

In this work, 48 prosodic and 16 formant frequency features were extracted, which were shown to be the most important factors in affect classification [Song *et al.* 2004; Zeng *et al.* 2005]. The extracted prosodic features include: max, min, mean, median of Pitch (Energy); mean, median of Pitch (Energy) rising/ falling slopes; max, mean, median duration of Pitch (Energy) rising/ falling slopes; mean, median value of Pitch (Energy) plateau at maxima/ minima; max, mean, median duration of Pitch (Energy) plateau at maxima/ minima.

$$\textit{If } |P(x)'-0|<\varepsilon \;\;\&\&\;\; P(x)''>0 \textit{, then } x \in \textit{a plateau at minima}$$
$$\textit{Else if } |P(x)'-0|<\varepsilon \;\;\&\&\;\; P(x)''<0 \textit{, then } x \in \textit{a plateau at maxima}$$

Where $P(x)$ is the Pitch (Energy) value of point x, $P(x)'$ is the first derivative and $P(x)''$ is the second.

Statistical properties of formant frequency including max, min, mean, median of the first, second, third, and fourth formant were extracted [Ververidis *et al.* 2004]. The acoustic feature analysis tool Praat is used to extract the Pitch, Energy and Formant of speech data. All features are based on a speech sentence.

In the experiment for clean speech, speaker-independent and speaker-dependent emotion recognitions were both investigated within the same gender. On the other hand, in the experiment for noisy speech, speaker-dependent emotion recognition was investigated. 10-fold cross-validation method was adopted considering the confidence of recognition results. 90% speech data were used for training and 10% for validation. 64-dimensional vectors of all speech data were projected into the six-dimensional space using the ELE method mentioned above.

4.3 Emotion Recognition in Speech

SVM, a powerful tool for classification, was introduced to classify the six emotions in this experiment. It had originally been proposed for two-class classification. In this system, 15 one-to-one SVMs were combined into an MSVM (Multi-SVM), in which each SVM was used to distinguish one emotion from another. Final classification result was determined by all the SVMs with the majority rule. After the heavy tests of polynomial, radial basis function and linear kernels with different parameters, linear SVM (C=0.1) was selected for its acceptable performance and simplicity.

In the experiment mentioned above, k=10 nearest neighbors were searched in constructing the distance matrix M and embedding the testing data. The impact of different k on the system performance was also investigated. Distribution of recognition accuracy from

clean speech on different k is shown in Figure 3. From the curve, influences made by k on male model are similar to that of female model. In both models, k =10 makes an acceptable performance with relatively low computational cost.

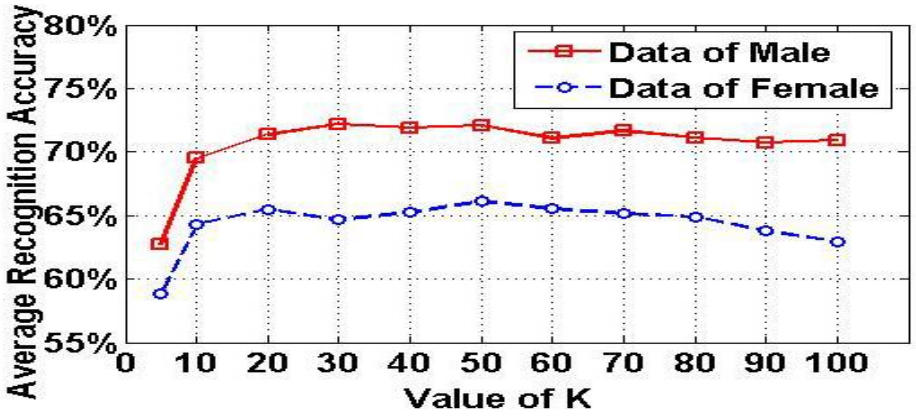
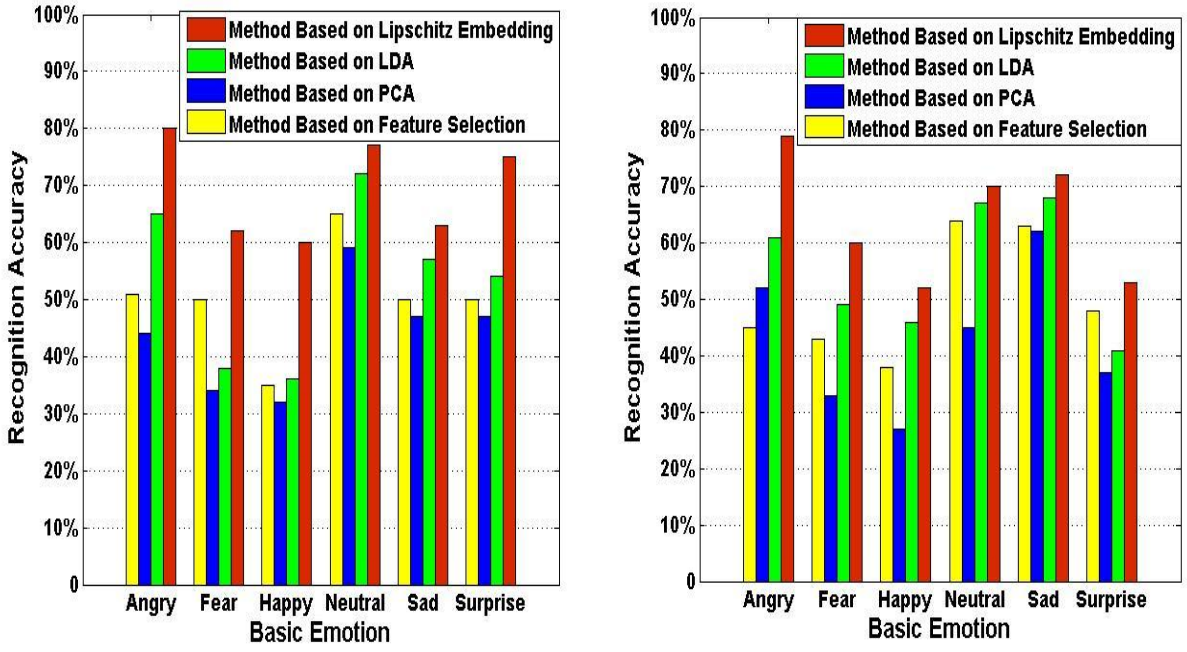


Figure 3. Distribution of recognition accuracy on different k

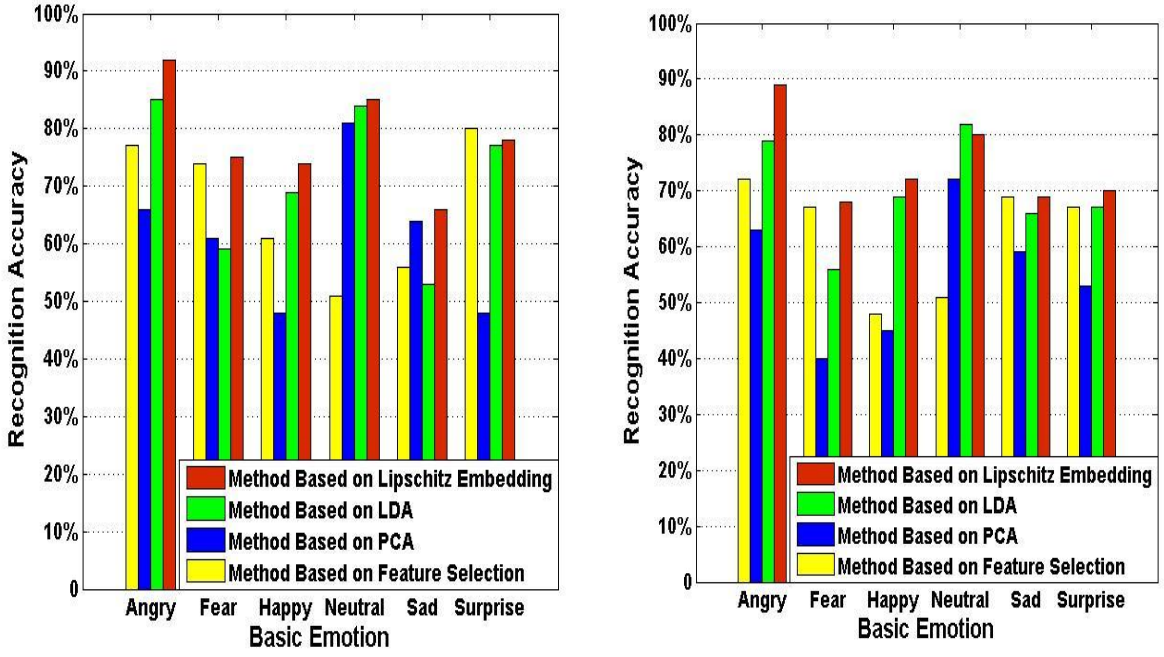
In order to evaluate the classification results of ELE, linear dimensionality reduction methods such as PCA, LDA and feature selection by SFS with an SVM classifier were also included for comparison. 64-dimensional features were projected into the six-dimensional space in every method. Figure 4 demonstrates the comparative performance of the four methods in speaker-independent emotion recognition. Speaker-dependent implementation results of the four methods are shown in Figure 5.

Due to acoustic variations that exist between different people, the average accuracy of the speaker-dependent emotion recognition (Figure 5) is about 10% higher than that of the speaker-independent (Figure 4). The classification rate of the male speaker is a little higher than the female, which probably indicates that women's facial expressions or body gestures convey more emotional information. In speaker-independent and speaker-dependent processes, the method based on ELE comes up with the best performance in almost all of the emotional situations. The relative improvement of the proposed method is 9%-26% in the speaker-independent system and 5%-20% when the system is speaker-dependent. While the classification rate of happiness is lower than other emotions in the speaker-independent system, the accuracy of happiness is comparable with the others in the speaker-dependent. What one can deduce from the results is that people express happiness in greatly varying manners.



(a) Male **(b) Female**

Figure 4. Speaker-independent performance comparison among the four methods.

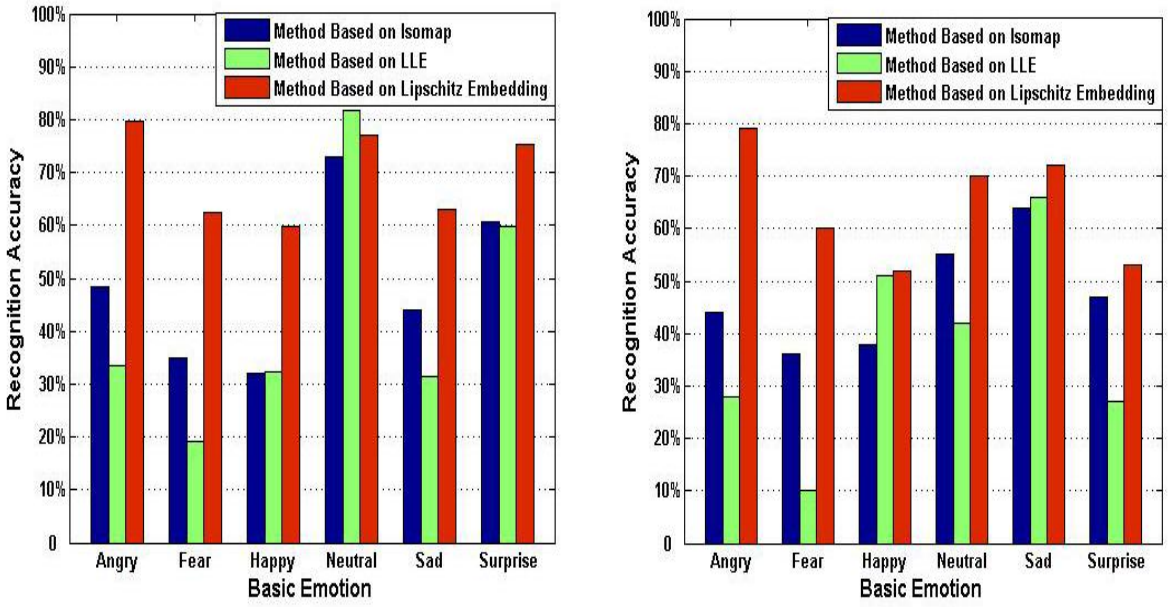


(a) Male **(b) Female**

Figure 5. Speaker-dependent performance comparison among the four methods.

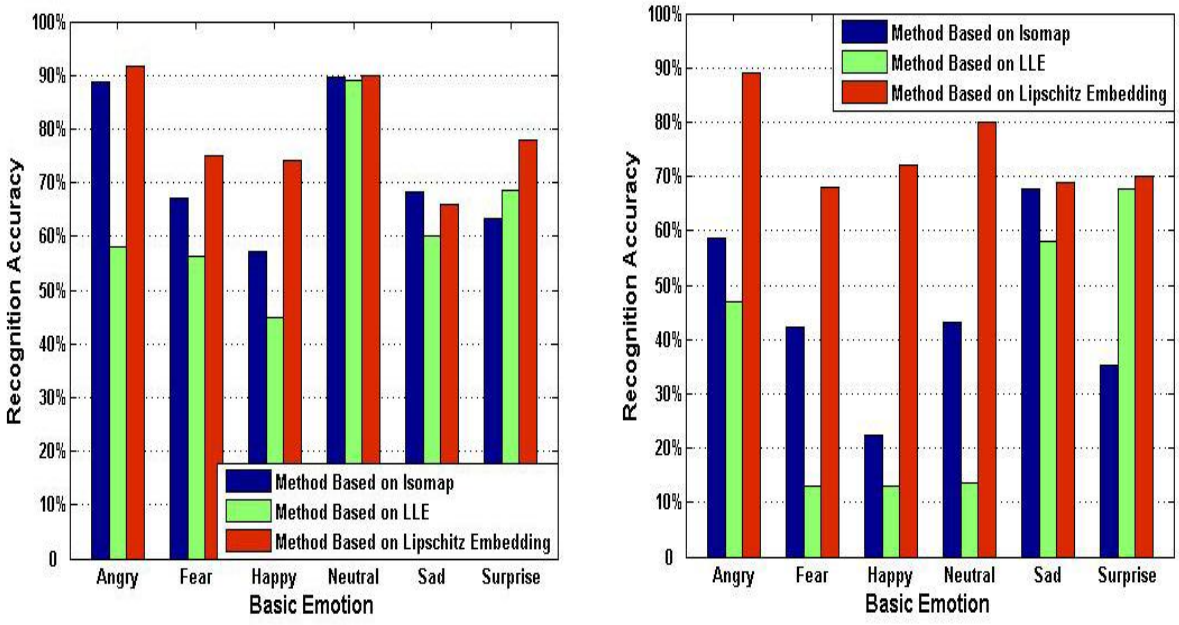
Considering the nonlinear submanifold that the ELE method involves, popular nonlinear dimensionality reduction approaches, such as Isomap and LLE, are implemented in this emotional speech recognition system for comparison. As mentioned before, Isomap and LLE only decide how to project the training data into a low dimensional manifold and leave the projection problem of novel testing data unsettled. However, in the emotion recognition system, all the training data and testing data should be embedded into low dimensional space. In the implementation of Isomap and LLE, the authors reconstruct the distance matrix M when facing the novel test data. Although it costs a lot of computation time, it will help attain Isomap and LLE's best performance. Comparison with those results gives one a solid evaluation of the proposed method.

Figure 6 and Figure 7 display the recognition accuracy of the six emotions in a speaker-independent and a speaker-dependent environment, respectively. From both figures, the method based on ELE still yields the best results in almost all of the emotional situations. In a speaker-independent environment, the proposed method outperforms the other two in the emotions angry and fear, especially. For the emotional speech recognition application, Isomap is more suitable than LLE. From Figure 6 and Figure 7, one can see that the recognition accuracy of Isomap is higher than LLE in most of the emotion states. Isomap is based on geodesic distance estimation and captures the global data structure when finding the low embeddings, while LLE focuses on preserving the local geometry of data points. ELE is somewhat similar to Isomap, which may explain why Lipschitz embedding and Isomap both outperform LLE in the experimental results. However, Isomap consumes more computation time than ELE. They both need the time-consuming operation of constructing the neighborhood graph, but the embedding step of Isomap is more complex. LLE conducts unbalanced performance when dealing with different basic emotions. For example, in Figure 6(b), LLE only attains 10% accuracy with the emotion fear, while it achieves about 65% with sad. The unbalanced recognition rate will greatly reduce the system's robustness. LLE gets a poor recognition rate for the female speaker in the speaker-dependent environment shown in Figure 7(b). Isomap and LLE behave differently between the male and the female in the speaker-dependent environment, but the performance of ELE seems stable. Comparing the results of Figure 6 and Figure 7 with those of Figures 4 and 5, nonlinear methods' performance may not be better than the linear methods', although they require more complicated computation. It is shown that the method of preserving the geometry of the data set is crucial in nonlinear manifold reduction approaches.



(a) Male **(b) Female**

Figure 6. Speaker-independent performance comparison between three nonlinear methods.

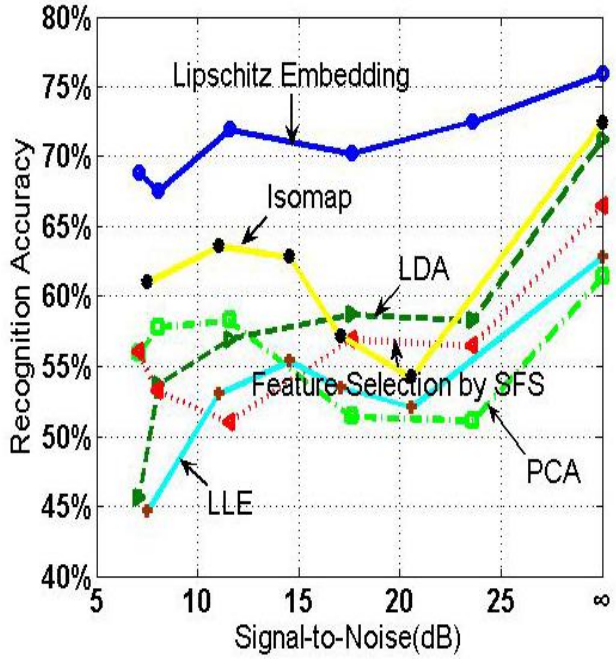


(a) Male **(b) Female**

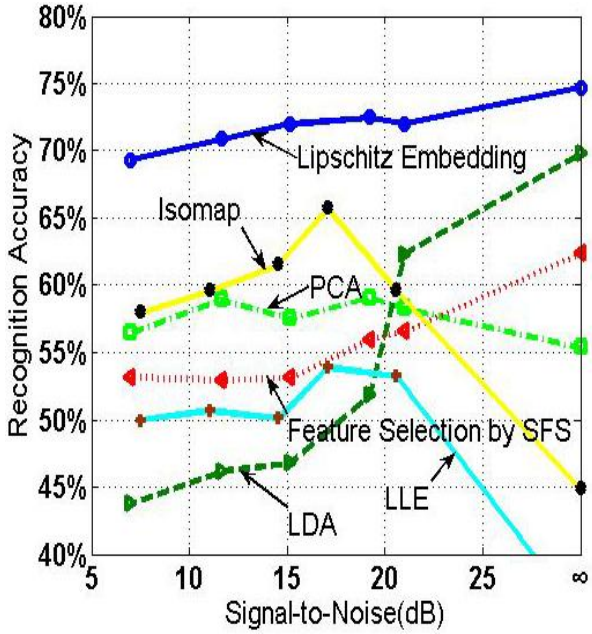
Figure 7. Speaker-dependent performance comparison between three nonlinear methods.

In order to test the perception constancy of the proposed approach, the classification performance of ELE on noisy speech is also investigated. Classical methods like PCA, LDA and feature selection by SFS with SVM were included for performance comparison. Nonlinear methods, Isomap and LLE, were also implemented. 64-dimensional features were projected into the six-dimensional space in every method. In many other applications, researchers tended to conduct noise reduction first for the noisy speech data. However, traditional noise reduction methods still face several challenges: the method using microphone array cannot avoid the problem of increasing the number of microphones; in the case of the spectral subtraction (SS) method, the musical tones arise from residual noise and processing delays also occur. With these considerations, the authors investigated emotion recognition from noisy speech directly, instead of conducting noise reduction. Since facial images with different poses and lighting directions were observed to make a smooth manifold, speech corrupted by noise may also be embedded into a low dimensional nonlinear manifold.

Figure 8 demonstrates the six methods' emotion recognition accuracy for clean speech and speech suppressed by Gaussian white noise. Performances with clean speech and speech corrupted by sinusoid noise are shown in Figure 9. Accuracies in both figures are the average recognition ratio of six emotions. From both figures, this system, based on Lipschitz embedding, shows outstanding performance with every SNR test data. Compared with the other methods, the accuracy of this method on Lipschitz embedding is stable both with speech corrupted by Gaussian white noise and with speech corrupted by sinusoid noise. Although there are differences among individuals, ELE is good at discovering the intrinsic geometry of the emotional speech manifold. The accuracy of LDA on clean speech is high, but drops quickly when noise increases. On the other hand, the accuracy of PCA can hardly be corrupted by louder noise, although its overall performance is poor. The Isomap Method also achieves acceptable accuracy in different experimental environments, except for the male speaker in speech corrupted by sinusoid noise. The performance of LLE is still disappointing. Keeping the local geometry by reconstructing from neighbors seems not to be appropriate for emotional speech recognition applications. From these figures, one can see an interesting phenomenon where the recognition accuracy of noisy speech is sometimes higher than that of clean speech. Features used to distinguish the different emotion states are strengthened by mild noise.

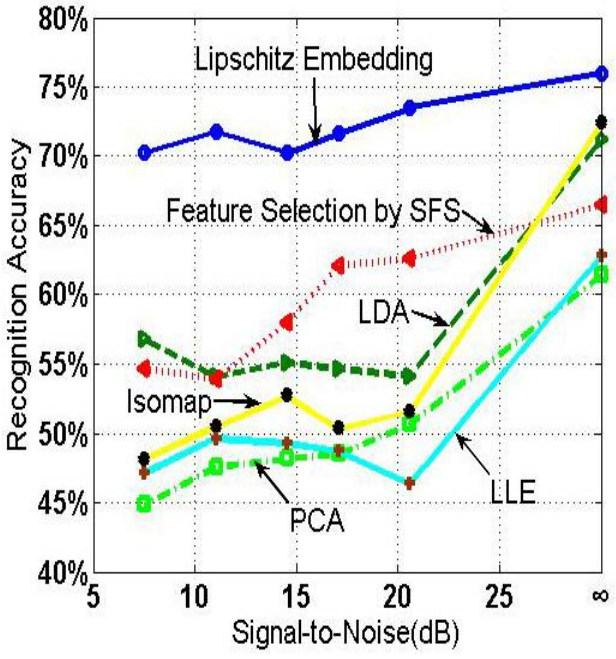


(a) Male

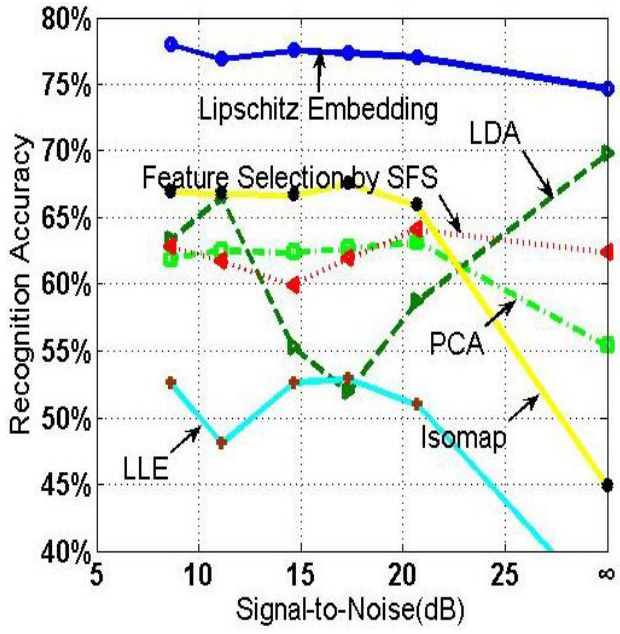


(b) Female

Figure 8. Performance comparison between linear and nonlinear methods on speech corrupted by Gaussian white noise. ∞ in the x-axis represents clean speech signal.



(a) Male



(b) Female

Figure 9. Performance comparison between linear and nonlinear methods on speech corrupted by sinusoid noise. ∞ in the x-axis represents clean speech signal.

5. Conclusion and Future Work

In this paper, the authors proposed an emotional speech recognition system based on a nonlinear manifold. Method ELE was presented to discover the intrinsic geometry of emotional speech including clean and noisy utterances. Compared with traditional approaches, including linear and nonlinear dimensionality reduction methods, this method came up with the best performance when dealing with almost all of the basic emotions in both speaker-independent and speaker-dependent processes. Even in a noisy environment, the performance of ELE was outstanding compared with the other methods and robust when different kinds of noise increase. Although LDA and Isomap also achieved plausible recognition results in the experiments, the proposed method balanced the classification rate in each emotion, which both of them lacked. The time consumption of Isomap was also higher than the proposed method. As another nonlinear method, LLE showed poor performance, meaning that preserving the intrinsic geometry of data corpus was vital. The key idea of the proposed method is to take the multiple classes of input patterns into consideration. Experimental results show that this idea is successful in emotional speech recognition applications.

With the method based on Lipschitz embedding, the average recognition accuracy of the female speaker is 5% lower than that of the male. The underlying reason should be investigated in detail and a robust algorithm is expected. Besides, the essential reason explaining the phenomenon that the accuracy of noisy speech exceeds clean speech should be investigated. In order to achieve better performance, improvement will be made to the proposed method and multi-modal emotion recognition will be included in future work.

ACKNOWLEDGEMENT

The work was partly supported by National Natural Science Foundation of China (60575032) and the 863 program (20060101Z1037). And the authors thank Cheng Jin and Can Wang for their generous help in the experiment and paper.

Reference

Bourgain, J., "On lipschitz embedding of finite metric spaces in hilbert space, " *Israel J. Math.*, 52(1-2), 1985, pp. 46-52.

Chang, Y., C. Hu, and M. Turk, "Probabilistic expression analysis on manifolds," In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004, Washington, DC, America, vol. 2, pp. 520-527.

Chuang, Z.J., and C. H. Wu, "Emotion recognition using acoustic features and textual content," In *Proceedings of IEEE International Conference on Multimedia and Expo*, 2004, Taipei, Taiwan, vol. 1, pp. 53-56.

Duchene, J., and S. Leclercq, "An optimal transformation for discriminant principal component analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 10(6), 1988, pp. 978-983.

Go, H., K. Kwak, D. Lee, and M. Chun, "Emotion recognition from the facial image and speech signal," In *proceedings of SICE 2003 Annual Conference*, 2003, Fukui, Japan, vol. 3, pp. 2890-2895.

Jain, V., and L. K. Saul, "Exploratory analysis and visualization of speech and music by locally linear embedding," In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, Montreal, Canada, vol. 3, pp. 984-987.

Johnson W., and J. Lindenstrauss, "Extension of lipschitz mapping into a hilbert space," *Contemporary Math.*, vol. 26, 1984, pp. 189-206.

Lee, C.M., S. S. Narayanan, and R. Pieraccini, "Classifying emotions in human-machine spoken dialogs," In *Proceedings of IEEE International Conference on Multimedia and Expo*, 2002, Lusanne, Switzerland, vol. 1, pp. 737-740.

Picard, R., *Affective computing*, The MIT Press, Cambridge, MA, 1997.

Roweis, S., and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, 2000, pp. 2323-2326.

Song, M., J. Bu, C. Chen, and N. Li, "Audio-visual based emotion recognition - a new approach," In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2004, Washington, DC, America, vol. 2, pp. 1020-1025.

Tenenbaum, J.B., V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, 2002, pp. 2319-2323.

Togneri, R., M. D. Alder, and Y. Attikiouzel, "Dimension and structure of the speech space," *IEEE Proceedings on Communications, Speech and Vision*, 139(2), 1992, pp. 123-127.

Ververidis, D., C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification," In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, Montreal, Canada, vol. 1, pp. 593-596.

Zeng, Z., Z. Zhang, B. Pianfetti, J. Tu, and T. S. Huang, "Audio-visual affect recognition in activation-evaluation space," In P*roceedings of IEEE International Conference on Multimedia and Expo*, 2005, Amsterdam, Netherlands, pp. 828-831.

Computational Linguistics and Chinese Language Processing
Vol. 12, No. 1, March 2007, pp. 65-78
© The Association for Computational Linguistics and Chinese Language Processing

Emotion Recognition from Speech Using IG-Based Feature Compensation

Chung-Hsien Wu*, and Ze-Jing Chuang*

Abstract

This paper presents an approach to feature compensation for emotion recognition from speech signals. In this approach, the intonation groups (IGs) of the input speech signals are extracted first. The speech features in each selected intonation group are then extracted. With the assumption of linear mapping between feature spaces in different emotional states, a feature compensation approach is proposed to characterize feature space with better discriminability among emotional states. The compensation vector with respect to each emotional state is estimated using the Minimum Classification Error (MCE) algorithm. For the final emotional state decision, the compensated IG-based feature vectors are used to train the Gaussian Mixture Models (GMMs) and Continuous Support Vector Machine (CSVMs) for each emotional state. For GMMs, the emotional state with the GMM having the maximal likelihood ratio is determined as the final output. For CSVMs, the emotional state is determined according to the probability outputs from the CSVMs. The kernel function in CSVM is experimentally decided as a Radial basis function. A comparison in the experiments shows that the proposed IG-based feature compensation can obtain encouraging performance for emotion recognition.

Keywords: Emotional Speech, Emotion Recognition, Intonation Group, Feature Compensation

1. Introduction

Human-machine interface technology has been investigated for several decades. Recent research has put more emphasis on the recognition of nonverbal information, especially on the topic of emotion reaction. Scientists have found that emotional skills can be an important component of intelligence, especially for human-human communication. Although human-computer interaction is different from human-human communication, some theories

* Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, ROC
E-mail: {chwu, bala}@csie.ncku.edu.tw

[Received September 8, 2006; Revised September 20, 2006; Accepted October 10, 2006]

have shown that human-computer interaction essentially follows the basics of human-human interaction [Reeves *et al*. 1996; Picard 1997; Cowie *et al*. 2001]. Scientists have found that emotion technology can be an important component in artificial intelligence, especially for human-human communication [Salovey *et al*. 1990]. Although the human-computer interaction is different from human-human communication, some theories show that human-computer interaction basically follows the fundamental forms of human-human interaction [Reeves *et al*. 1996]. In this study, an emotion recognition approach from speech signals is proposed. This method consists of the definition and extraction of intonation groups (IGs), IG-based feature extraction, and feature compensation.

In past years, many researchers have paid attention to emotion recognition via speech signals. Several important recognition models have been applied to the emotion recognition task, such as Neural Network (NN) [Bhatti *et al*. 2004], Hidden Markov Model (HMM) [Inanoglu *et al*. 2005], Support Vector Machine (SVM) [Kwon *et al*. 2003; Chuang *et al*. 2004], and others [Subasic *et al*. 2001; Wu *et al*. 2006; Silva *et al*. 2000]. Besides the generally used prosodic and acoustic features, some special features are also applied for this task, such as TEO-based features [Rahurkar *et al*. 2003]. Although lots of features and recognition models have been tested in these works, large overlaps between the feature spaces for different emotional states is rarely considered. Besides, the pre-trained emotion recognition model is highly speaker-dependent [Chuang *et al.* 2004; Wu *et al*. 2004].

To solve the above questions, this paper proposes an approach to emotion recognition based on feature compensation. The block diagram of the approach is shown in Figure 1. The feature extraction process is shared by the training and testing phase and is divided into two steps: intonation group (IG) identification and IG-based feature extraction. In order to identify the most significant segment, the intonation groups (IGs) of the input speech signals are first extracted. Following the feature extraction process [Deng *et al*. 2003], the prosodic feature sets are estimated for the IG segments. Then, in training phase, the extracted feature vectors are applied for compensation vector estimation. All the feature vectors compensated by compensation vectors are modeled by a Gaussian Mixture Model (GMM). Finally, the minimum classification error (MCE) training method [Wu *et al*. 2002] iteratively estimates all the model parameters. As a comparison, the compensated vectors are also used to train the Continuous Support Vector Machine (CSVM) model. In the testing phase, the extracted feature vectors are directly compensated using the compensation vectors. Then, the final emotional state is decided using the CSVM model.

The rest of the paper is organized as follows. Section 2 describes the definition of Intonation Group and the extraction of the prosodic features. Then the feature compensation technique and MCE training is provided in Section 3. The model description of CSVM is shown in Section 4. Finally, experimental results and conclusions are drawn in Section 5 and 6,

respectively.

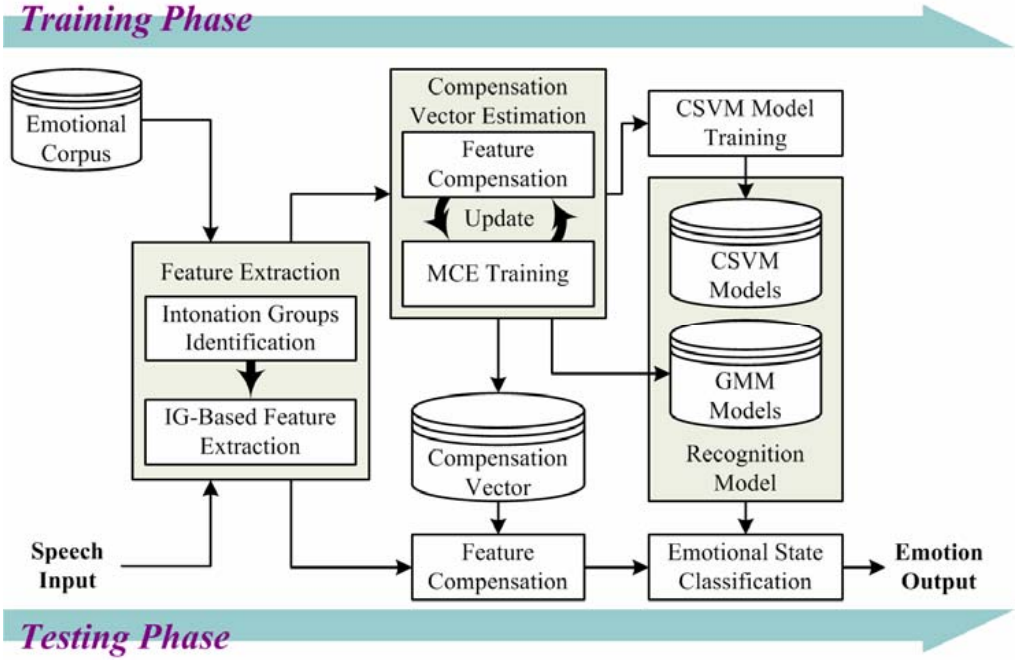


Figure 1. Block diagram of the proposed emotion recognition approach

2. IG-Based Feature Extraction

2.1 Intonation Group Extraction

The intonation group, also known as breath-groups, tone-groups, or intonation phrases, is usually defined as the segment of an utterance between two pauses.

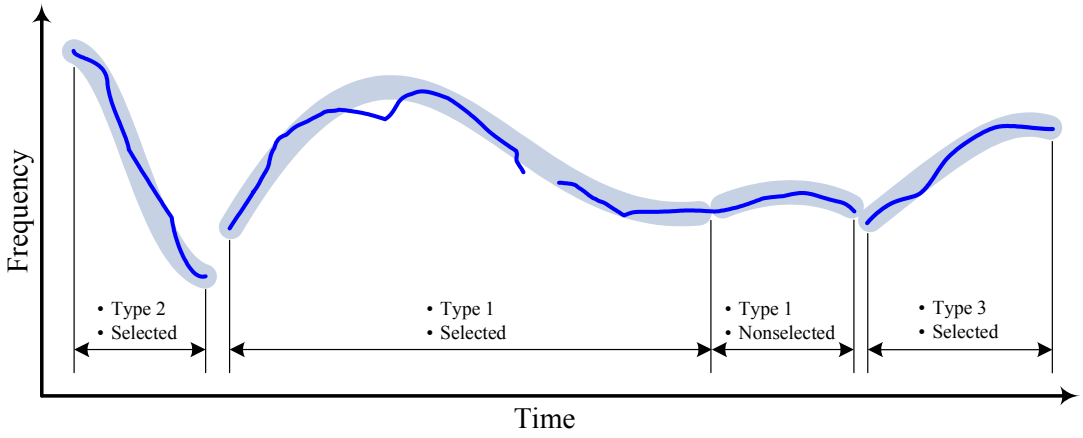


Figure 2. An illustration of the definition and extraction of Intonation Groups. Four IGs are extracted from the smoothed pitch contour (the gray-thick line), but only three IGs (the first, second, and forth IGs) are selected for feature extraction

As shown in Figure 2, the intonation group is identified by analyzing the smoothed pitch contour (the gray-thick line in Figure 2). Three types of smoothed pitch contour patterns are defined as the intonation group:

- **Type 1**: a complete pitch segment that starts from the point of a pitch rise to the point of the next pitch rise,
- **Type 2**: a monotonically decreasing pitch segment,
- **Type 3**: a monotonically increasing pitch segment.

For all identified IG segments, only those IGs that match the following criterion are selected for feature extraction:

- the complete IGs with the largest pitch range or duration,
- the monotonically decreasing or increasing IGs with the largest pitch range or duration,
- the monotonically decreasing or increasing IGs at the start or end of a sentence.

In Figure 2, the numbers before the slash symbol indicate the type of IG, and the symbol S and NS indicate "Selected" and "Not-Selected" IGs, respectively. Although there are four IGs extracted, only three IGs are selected for feature extraction.

2.2 IG-Based Feature Extraction

Emotional state can be characterized by many speech features, such as pitch, energy, or duration [Ververidis *et al*. 2005]. In this paper, the authors use the following 64 prosodic features as the input features for emotion recognition:

- Speaking rate and relative duration (2 values). The relative duration is normalized with respect to the length of the input sentence.
- Pause number and relative pause duration (2 values). The relative duration is normalized with respect to the length of intonation group. The same definition of relative position and relative duration are made in the following features.
- Mean and standard deviation of pitch, energy, zero-crossing-rate, and F1 values (8 values).
- Mean and standard deviation of jitter (for pitch) and shimmer (for energy) (4 values).
- Maximum and minimum of pitch, energy, zero-crossing-rate, and F1 values (8 values).
- Relative positions at which the maximal pitch, energy, zero-crossing-rate, and F1 value occur (4 values).
- Relative positions at which the minimal pitch, energy, zero-crossing-rate, and F1 value occur (4 values).
- Fourth-order Legendre parameters of pitch, energy, zero-crossing-rate, and F1 contours of the whole sentence (16 values).

- Fourth-order Legendre parameters of pitch, energy, zero-crossing-rate, and F1 contours inside the "significant segment," which is the segment during the positions of maximum and minimum values (16 values).

The definitions of jitter and shimmer are given in [Levity *et al*. 2001]. Jitter is a variation of individual cycle lengths in pitch-period measurement, while shimmer is the measure for energy values. The calculation of the jitter contour is shown as:

$$J_i = \frac{\sum_{u=0}^{m} f_i - ug_u}{f_i},\ g = \frac{1}{4}\{-1,3,-3,1\}, \tag{1}$$

where variable f_i indicates the i-th pitch value. The mean and standard deviation of jitter are calculated by equations 2 and 3, respectively.

$$J_{mean} = \frac{\sum_{J_i<J_\theta} J_i}{\#_{J_i<J_\theta} J_i},$$
$$J_{dev} = \sqrt{\frac{\sum_{J_i<J_\theta} \left(J_i - J_{mean}\right)^2}{\#_{J_i<J_\theta} J_i}}. \tag{2}$$

where variable J_θ is a threshold of jitter used to avoid the outlier noise.

3. Compensation Vector Estimation Using MCE

The goal of feature compensation is to move the feature space of an emotional state to a feature space more discriminative to other emotional states. Given a sequence of training data $\mathbf{X}^e = \left\{x_n^e\right\}_{n=1}^N$, where x_n^e indicates the *n*-th feature vector that belongs to emotional state E_e. The feature vector extracted for each intonation group contains the prosodic features mentioned above. With the assumption of linear mapping between feature spaces in different emotional states, the vector compensation function is defined as:

$$\tilde{x}_n^{e\to f} = x_n^e + p\left(E_e \left| x_n^e \right.\right) r_{e\to f}, \tag{3}$$

where $r_{e\to f}$ is a compensation vector of emotional state E_e with respect to the reference emotional state E_f. The conditional probability of the emotional state E_e given the input feature vector x_n^e is estimated as:

$$p\left(E_e \left| x_n^e \right.\right) = \frac{p\left(x_n^e \left| E_e \right.\right) p\left(E_e\right)}{\sum_i p\left(x_n^e \left| E_i \right.\right) p\left(E_i\right)}. \tag{4}$$

Minimum classification error (MCE) training based on the generalized probabilistic descent (GPD) method is applied in this study. The authors assume that the probability of a mapped feature vector $\tilde{x}_n^{e\to f}$ given an emotional state E_c follows the distribution of a mixture of Gaussian density function:

$$g_c\left(\tilde{x}_n^{e\to f}\right) = \sum_m w_m^c \cdot N\left(\tilde{x}_n^{e\to f}; \mu_m^c, \delta_m^c\right), \tag{5}$$

where $N\left(\cdot; \mu_m^c, \delta_m^c\right)$ denotes the normal distribution with mean μ_m^c, and diagonal covariance matrix δ_m^c, and w_m^c is the mixture weight. To estimate the mapping coefficients and GMM parameters jointly by MCE training, the misclassification measure is defined as:

$$D_e \equiv D(\mathbf{X}_e) = -g_e(\mathbf{X}_e) + \frac{1}{\eta}\log\left[\frac{1}{\mathbf{C}-1}\sum_{c\neq e}\exp\left(\eta \cdot g_c(\mathbf{X}_e)\right)\right], \tag{6}$$

where $\mathbf{X}_e$ denotes a set of data compensated from the emotional state E_e, $\mathbf{X}_e = \left\{\tilde{x}_n^{e\to f}\right\}_{f\neq e}$, $\mathbf{C}$ is the number of emotional state, and η is a penalty factor. The function $g_c\left(\mathbf{X}_e\right)$ is the average likelihood estimated by the GMM of the emotional state $\boldsymbol{E}_c$ given $\mathbf{X}_e$. Based on the GPD iterative theory, the parameters will approximate the global optimization using the iterative equation:

$$\Theta_{t+1} = \Theta_t - \varepsilon \cdot \nabla l, \tag{7}$$

The loss function is defined as a sigmoid function of misclassification measure. And the gradient of loss function ∇l is the partial differential to the updated parameter. Using chain rule, the gradient of loss function can be divided into three components. The first component can be derived to a closed form $a \cdot l_e \cdot \left(1 - l_e\right)$, and the second component is assumed as:

$$\frac{\partial D_e}{\partial g_c} = \begin{cases} -1 & , e = c \\ 1 & , e \neq c \end{cases}. \tag{8}$$

Since there are four different parameters needing to be updated, the last component of the gradient with respect to each parameter is obtained as:

$$\frac{\partial g_e}{\partial r_{e\to f}} = -\mathbf{A}\sum_n\sum_m\left[\frac{w_m^e\left(\tilde{x}_n^{e\to f} - \mu_m^e\right)p\left(E_e \mid x_n^e\right)}{\left(\delta_n^e\right)^2} N\left(x_n^e; \mu_m^e, \delta_n^e\right)\right], \tag{9}$$

$$\frac{\partial g_e}{\partial w_m^e} = \mathbf{A}\sum_n \sum_r \mathbf{B}\,, \tag{10}$$

$$\frac{\partial g_e}{\partial \mu_m^e} = \mathbf{A}\sum_n \sum_r \left[w_m^e \left(\tilde{x}_n^{e\to f} - \mu_m^e \right) \left(\delta_m^e \right)^{-2} \mathbf{B} \right], \tag{11}$$

$$\frac{\partial g_e}{\partial \delta_m^e} = \mathbf{A}\sum_n \sum_e \left[w_m^e \left(\left(\tilde{x}_n^{e\to f} - \mu_m^e \right)^2 - \left(v_m^e \right)^2 \right) \left(v_m^e \right)^{-3} \mathbf{B} \right]. \tag{12}$$

where

$$\mathbf{A} = \frac{1}{\mathbf{N}(\mathbf{C}-1)} \quad , \quad \mathbf{B} = N\left(\tilde{x}_n^{e\to f}; \mu_m^e, \delta_m^e \right)$$

Given an input feature vector *y*, the recognized emotional state E_e^* is determined according to the following equation:

$$E_e^* = \arg\max_e \left[\frac{\sum_{i\neq e} g_e\left(y + p\left(E_e \middle| y \right) r_{e\to i} \right)}{\sum_{j\neq e} g_j\left(y + p\left(E_j \middle| y \right) r_{j\to e} \right)} \right]. \tag{13}$$

4. Emotion recognition using CSVM models

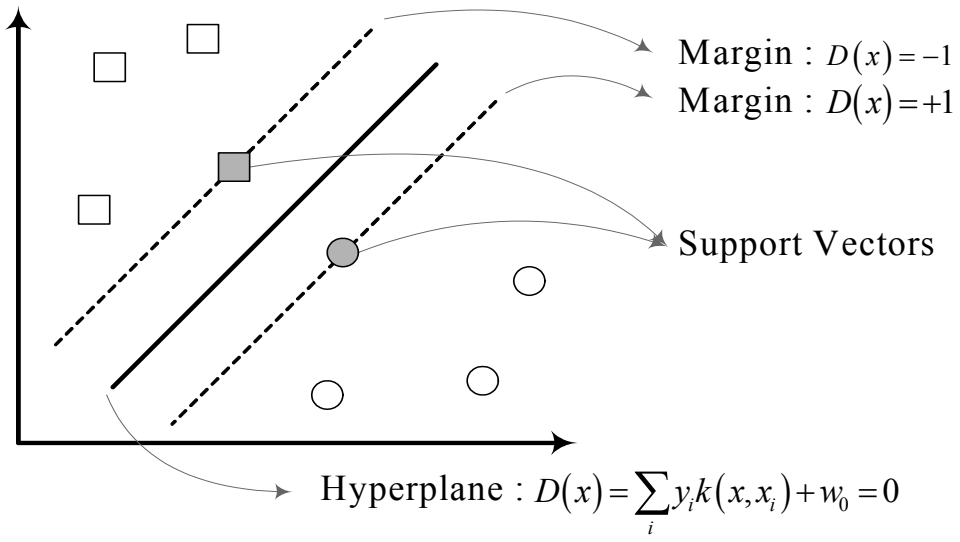


Figure 3. An illustration of SVM. The vectors on the margins are the so-called "Support Vectors"

The SVM has been widely applied in many research areas, such as data mining, pattern recognition, linear regression, and data clustering. Given a set of data belonging to two classes, the basic idea of SVM is to find a hyperplane that can completely distinguish two different classes. The illustration of SVM model is shown in Figure 3. The hyperplane is decided by the maximal margin of two classes, and the samples that lie in the margin are called "support vectors." The equation of the hyperplane is described as:

$$D(x) = \sum_{i=1}^{N} y_i k(x \cdot x_i) + w_0 \ , \tag{14}$$

where $k(x \cdot x_i)$ is kernel function. Traditional SVMs can construct a hard decision boundary with no probability output. In this study, SVMs with continuous probability output are proposed. Given the test sample *x'*, the probability that *x'* belongs to class *c* is $P(class_c|x')$. This value is estimated based on the following factors:

- the distance between the test input and the hyperplane,

$$R = \frac{D(x')/\|w\|}{1/\|w\|} = D(x') \quad ; \tag{15}$$

- the distance from the class centroid to the hyperplane,

$$R' = \frac{R}{D(\bar{x})} = \frac{D(x')}{D(\bar{x})} \quad ; \tag{16}$$

  where $\bar{x}$ is the centroid of the training data in a class;

- the classification confidence of the class P_c, which is defined as the ratio of correctly recognized sentences number to the total sentence number.

Finally, the output probability is defined as follows, in accordance with the above factors:

$$P(class_c|x') = \frac{P_c}{1+\exp(1-R')} = \frac{P_c}{1+\exp\left(1-\frac{D(x')}{D(\bar{x})}\right)} \quad . \tag{17}$$

The CSVM model with the highest probability determines the emotion output.

5. Experimental Results

In this experiment four kinds of emotional states: Neutral, Happy, Angry, and Sad were adopted. The emotional speech corpus was collected in an 8 KHz sampling rate and a 16-bit resolution. 40 sentences for each emotional state were recorded by 8 volunteers.

In addition to the proposed prosodic features, the researchers also evaluated the recognition rate for Mel-Frequency Cepstrum Coefficient (MFCC) features, which are generally used in a speech recognition task. Both GMMs and CSVMs are applied in the experiments to compare with the baseline system, which is a GMM emotion recognition system without any preprocessing before feature extraction.

5.1 Mixture Number Determination of GMM

The number of mixtures in the GMM is first determined for each emotional state. Assuming that the number of mixtures is greater than 3, the average likelihood of all training data given the GMM with different mixture number is calculated. Figure 4 shows the plot of GMM likelihoods using both prosodic and MFCC features. Accordingly, the numbers of mixtures using prosodic features is set to 6, 5, 12, and 9 for neutral, angry, happy, and sad emotion, respectively. To evaluate MFCC features, the number of mixtures was also evaluated using the same method. The contours of the likelihood using MFCC features is shown in the lower part of Figure 4, and the mixture numbers for neutral, angry, happy, and sad emotions are set to 13, 11, 18, and 20, respectively.

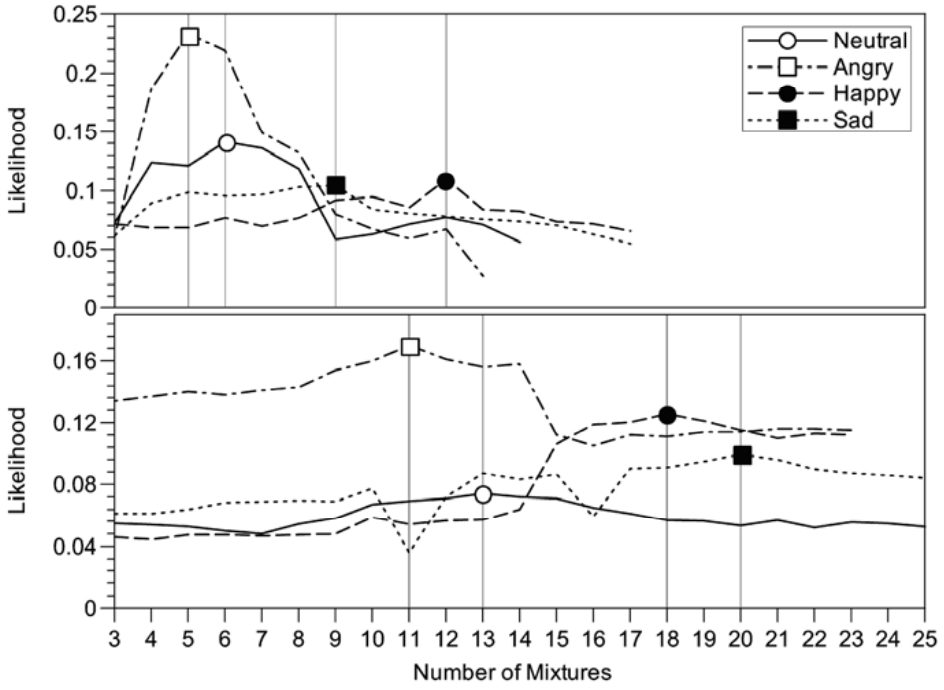


Figure 4. Likelihood contours of GMMs with increasing mixture number. The upper part is the plot using prosodic features, and the lower part is the plot using MFCC features

5.2 Experiments on SVM Kernel Function

The kernel function defined in CSVM model is used to transfer a vector in original vector space to a new space with higher dimension. There are several popularly used kernel functions:

- Simple dot

$$k(x, y) = x \cdot y \tag{18}$$

- Vovk's polynomial

$$k(x, y) = (x \cdot y + 1)^p \tag{19}$$

- Radial basis function

$$k(x, y) = \exp\left(-\|x - y\|^2 / 2\delta^2\right) \tag{20}$$

- Sigmoid kernel

$$k(x, y) = \tanh\left(k(x, y) - \Theta\right) \tag{21}$$

In order to select the most appropriate kernel function, a primary test of emotion recognition using CSVM model with different kernel functions was applied. The primary test used both prosodic and MFCC features with feature compensation and intonation group. The result of emotion recognition is shown in Table 1. It is obvious that the Radial basis function is the most suitable kernel function for this test.

Table 1. The primary test for different kernel functions.

	Prosodic Feature	**MFCC**
Simple dot	59.00%	63.48%
Vovk's polynomial	60.83%	90.12%
Radial basis function	80.72%	95.12%
Sigmoid kernel	73.50%	81.46%

5.3 Experiments on Emotion Recognition using GMM

Table 2 shows the results for emotion recognition using GMM, including the proposed approach and the baseline system. In the first column, the abbreviations **FC**, **In**, **OO**, and **OC** indicate the methods using feature compensation, the results from **In**side, **O**utside-**O**pen, and **O**utside-**C**losed tests, respectively. The first row in Table 2 shows four kinds of speech features from left to right: frame-based prosodic feature, frame-based MFCC feature, IG-based prosodic feature, and IG-based MFCC feature.

Table 2. Emotion recognition results using GMMs. The abbreviation FC indicates the method using Feature Compensation. The abbreviation In, OO, and OC indicate the results from Inside, Outside-Open, and Outside-Closed tests, respectively

	Prosodic Feature	MFCC	Prosodic Feature+IG	MFCC+IG
Without FC (In)	74.33%	99.96%	76.32%	99.24%
Without FC (OO)	49.78%	35.15%	51.07%	35.01%
Without FC (OC)	55.95%	37.22%	59.90%	42.13%
With FC (In)	80.72%	95.12%	83.94%	91.32%
With FC (OO)	55.19%	41.27%	60.13%	49.10%
With FC (OC)	61.03%	41.03%	67.52%	52.86%

Although MFCC feature outperformed prosodic features in the inside test, prosodic features achieved better performance in both outside-open and outside-closed tests. The reason for this result is that the MFCC features contain a considerable amount of information from the speech content and the speaker. Emotional state modeling using MFCC features is highly related to speech content and speaker. Therefore, the GMM model can better classify the emotional states of the trained MFCC features, but cannot well characterize the unseen features in the outside test. In the proposed approach, MFCC features retain their higher recognition rate in the inside test, and the prosodic features obtain the best overall performance. From the above experiments, an increase in recognition rate for the approaches with IG-based feature extraction is about 5% to 10% compared to those without IG-based feature extraction. Furthermore, an improvement of 10% in recognition rate for the approaches with feature compensation is obtained compared to those without feature compensation.

5.4 Experiments on Emotion Recognition using CSVM

Table 3. Emotion recognition results using CSVMs.

	Prosodic Feature	MFCC	Prosodic Feature+IG	MFCC+IG
Without FC (In)	78.23%	98.12%	81.19%	99.15%
Without FC (OO)	50.83%	32.76%	53.79%	34.91%
Without FC (OC)	58.10%	38.48%	60.94%	40.33%
With FC (In)	83.10%	92.71%	86.08%	91.73%
With FC (OO)	57.36%	40.08%	62.12%	47.60%
With FC (OC)	62.55%	41.98%	68.00%	53.25%

Table 3 shows the results for emotion recognition using CSVM. General speaking, the results of emotion recognition using CSVM are similar to the results using GMM. The proposed feature extraction method, IG-based prosodic feature with feature compensation, obtains the

best overall performance. As seen in the result shown in Table 2, the MFCC feature attains a better recognition rate than the prosodic feature in the inside test, but worse in the outside test. The difference between using CSVM and GMM is the suitability of CSVM in a data sparse situation. Since CSVM constructs the classification hyperplane using only few support vectors, it attains better classification results than other classification methods when the training corpus is insufficient. In the proposed method, the MFCC feature is extracted frame by frame, while the prosodic feature is extracted by a single sentence or IG segment. It is obvious that, under the same number of training corpora, the number of MFCC features is larger than the number of prosodic features. Accordingly, with the prosodic features, most results using CSVM are better than that using GMM.

6. Conclusion

In this paper, an approach to emotion recognition from speech signals is proposed. In order to obtain crucial features, the IG-based feature extraction method is used. After feature extraction, the feature vector compensation approach and MCE training method are applied to increase the discriminability among emotional states. The experiments show that it is useful to integrate IG-based feature extraction and feature compensation to emotion recognition. The result of emotion recognition using the proposed approaches with GMM is 83.94% for an inside test and 60.13% for an outside-open test, and the result with CSVM is 86.08% for an inside and 62.12% for an outside-open test. This result shows that the CSVM classification model is more suitable than GMM when performing the emotion recognition task. The authors also demonstrate that the prosodic feature is more suitable for emotion recognition than the acoustic MFCC features in speaker-independent task.

The future work of this research is to improve the recognition accuracy for outside data. Though the feature compensation is useful for emotion recognition, the compensation vector is still speaker-dependent. An adaptation method will be useful to adapt compensation vectors for emotional speech with different speaking styles.

References

Bhatti, M.W., Y. Wang, and L. Guan, "A neural network approach for human emotion recognition in speech," In *Proceedings of the 2004 IEEE International Symposium on Circuits and Systems*, Vancouver, Canada, pp. 1811-1184.

Chuang, Z.J., and C.H. Wu, "Multi-Modal Emotion Recognition from Speech and Text," *International Journal of Computational Linguistics and Chinese Language Processing*, 9(2), 2004, pp. 45-62.

Chuang, Z.J., and C.H. Wu, "Emotion Recognition using Acoustic Features and Textual Content," In *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo*, 2004, Taipei, Taiwan, pp. 53-56.

Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, 18(1), 2001, pp. 32-80.

Deng, L., J. Droppo, and A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Transactions on Speech and Audio*, 11(6), 2003, pp. 568-580.

Inanoglu, Z., and R. Caneel, "Emotive alert: HMM-based emotion detection in voicemail messages," In *Proceedings of IEEE Intelligent User Interfaces*, 2005, San Diego, California, USA, pp. 251-253.

Kwon, O.W., K. Chan, J. Hao, and T.W. Lee, "Emotion Recognition by Speech Signals," In *Proceedings of the 8th European Conference on Speech Communication and Technology*, 2003, Geneva, Switzerland, pp. 125-128.

Levity, M., R. Huberz, A. Batlinery, and E. Noeth, "Use of prosodic speech characteristics for automated detection of alcohol intoxication," In *Proceedings of the Workshop on Prosody and Speech Recognition 2001*, Red Bank, NJ, pp. 19-22.

Picard , R.W., *Affective Computing*, Cambridge, MA: MIT Press, 1997.

Rahurkar, M.A., and J.H.L. Hansen, "Frequency Distribution Based Weighted Sub-Band Approach for Classification of Emotional/Stressful Content in Speech," In *Proceedings of 8th European Conference on Speech Communication and Technology*, 2003, Geneva, Switzerland, pp. 721-724.

Reeves, B., and C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*, University of Chicago Press, 1996.

Reeves, B., and C. Nass, *The Media Equation*, Center for the Study of Language and Information, 1996.

Salovey, P., and J. Mayer, "Emotional Intelligence," *Imagination, Cognition and Personality*," 9(3), 1990, pp. 185-211.

Silva, L.C.D., and P.C. Ng, "Bimodal Emotion Recognition," In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, Grenoble, France, pp. 332-335.

Subasic, P., and A. Huettner, "Affect Analysis of Text Using Fussy Semantic Typing," *IEEE Transactions on Fussy System*, 9(4), 2001, pp. 483-496.

Ververidis, D., C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification," In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005, Montreal, Montreal, Canada, pp. 593-596.

Wu, J., and Q. Huo, "An environment compensated minimum classification error training approach and its evaluation on aurora2 database," In *Proceedings of 7th International Conference on Spoken Language*, 2002, Denver, Colorado, USA, pp. 453-456.

Wu, C.H. and Z.J. Chuang, "Intelligent Ear for Emotion Recognition: Multi-Modal Emotion Recognition via Acoustic Features, Semantic Contents and Facial Images," In *Proceedings of The 8th World Multi-Conference on Systemics, Cybernetics and Informatics*, 2004, Orlando, Florida, USA.

Wu, C.H., Z.J. Chuang, and Y.C. Lin, "Emotion Recognition from Text using Semantic Label and Separable Mixture Model," *ACM Transactions on Asian Language Information Processing*, 5(2), 2006, pp. 165-183.

Computational Linguistics and Chinese Language Processing
Vol. 12, No. 1, March 2007, pp. 79-90
© The Association for Computational Linguistics and Chinese Language Processing

Emotional Recognition Using a Compensation Transformation in Speech Signal

Cairong Zou[*] , Yan Zhao[+], Li Zhao [+], Wenming Zhen[+], and Yongqiang Bao[+]

Abstract

An effective method based on GMM is proposed in this paper for speech emotional recognition; a compensation transformation is introduced in the recognition stage to reduce the influence of variations in speech characteristics and noise. The extraction of emotional features includes the globe feature, time series structure feature, LPCC, MFCC and PLP. Five human emotions (happiness, angry, surprise, sadness and neutral) are investigated. The result shows that it can increase the recognition ratio more than normal GMM; the method in this paper is effective and robust.

Key words: Speech Emotional Recognition (SER), GMM, Emotion Recognition, Compensation Transformation

1. Introduction

One of the natural goals for research on speech signals is recognizing emotions of humans [Chen 1987; Oppenheim 1976; Cowie 2001]; it has gained growing amounts of interest over the last 20 years. A study conducted by Shirasawa *et al*. showed that SER could be made by ICA and attain an 87% average recognition ratio [Shirasawa 1997; Shirasawa 1999] Many studies have been conducted to investigate neural networks for SER. Chang-Hyun Park tried to recognize sequentially inputted data using DRNN in 2003[Park *et al*. 2003], Muhammad, W. B. obtained about 79% recognition rate using GRNN [Bhatti *et al*. 2004]. Aishah Abdul Razak achieved an average recognition rate of 62.35% using combination MLP [Razak *et al*. 2005]. Fuzzy rules are also introduced into SER such that an 84% rate has been achieved in recognizing anger and sadness [Austermann *et al*. 2005]. A number of studies in SER have

[*] Foshan University, Foshan, 528000, Guangdong, China
[+] Research Center of Learning Science, Southeast University, Nanjing, 210096, China
E-mail: zhaoli@seu.edu.cn

[Received August 11, 2006; Revised September 30, 2006; Accepted October 22, 2006]

also been done with the development of GMM/HMM [Rabiner 1989; Jiang *et al*. 2004; Lin *et al.* 2005]. However, in SER, the variations in speech characteristics, noise and individual differences always influence the recognition results. In addition, the methods above have always handled such problems in the preprocessing stage and have not been able to eliminate the influence effectively. Therefore, a valid solution has still not been proposed. In this paper a compensation transformation is introduced into an algorithm for GMM which operates in the recognition module. The experiments with five emotions (happiness, angry, neutral, surprise and sadness) show that the method in this paper is effective in emotional recognition.

2. Descriptions of Emotion and Selection of Emotion Speech Materials

Usually, emotions are classified into two main categories: basic emotions and derived emotions. Basic emotions, generally, can be found in all mammals. Derived emotions mean derivations from basic emotions. One viewpoint is that the basic emotions are composed by the basic mood. Due to different research backgrounds, different researchers have expressed different definitions of basic emotions. Some of the major definitions [Ortony *et al.* 1990] of the basic emotions are shown in Table 1.

Table 1. Researches about basic emotions definition

Researchers	definitions
Plutchik	Acceptance, joy, anger, anticipation, disgust, fear, sadness, surprise
Ekman/Friesen/ Ellsworth	Anger, disgust, fear, joy, sadness, surprise
James	Fear, grief, love, rage
Izard	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
Oatley/Johnson -Laird	Anger, disgust, anxiety, happiness, sadness
Panksepp	Expectancy, fear, rage, panic
Weiner/Graham	Happiness, sadness

The common emotion classification which was proposed by Plutchik is shown in Figure 1e. In this paper, the authors only recognize five kinds of emotion.

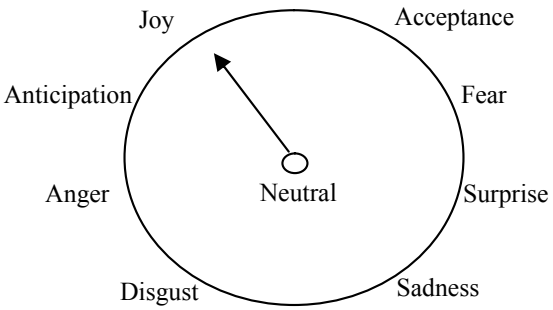


Figure 1. Emotion wheel

This is a relatively conservative view of what emotion is so special attention has been paid to emotional dimension space theory. Three major dimensions (valence, arousal, and control) [Cowie 2001] are used to describe emotions.

a. Valence: The clearest common element of emotional states is that the person is materially influenced by feelings that are valenced, *i.e*., they are centrally concerned with positive or negative evaluations of people or things or events.

b. Arousal: It has been proven that emotional states involve dispositions to act in certain ways. A basic way of reflecting that theme turns out to be surprisingly useful. States are simply rated in terms of the associated activation level, *i.e*., the strength of the person's disposition to take some action rather than none.

c. Control: Embodying in the initiative and the degree of control. For instance, contempt and fear are in different ends of the control dimension.

In this paper, two aspects have to be taken into consideration in the selection of emotional materials: 1. the sentence materials can't have any emotional tendency; 2. the materials should relate to five kinds of emotions (happiness, angry, surprise, sadness, and neutral). All recordings were carried out in a large, soundproof room with no echo interference using a high quality microphone, a SONY DAT recorder and a PC164 audio card at a sampling rate of l2KHZ with 16-bit resolution. Six speakers (three male and three female) who are good at acting spoke the sentences with happiness, anger, surprise and sadness, expressing each emotion three times. At the same time, the researchers made the speakers speak each sentence three times in a neutral way. In this way, 2430 sentences for experiments were compiled.

3. Feature Extraction

The emotional features of speech signals are always represented as the change of speech rhythm [Shigenaga 1999; Muraka 1998]. For example, when a man is in a rage, his speech rate, volume and tone will all get higher. Some characteristics of phonemes can also reflect the change of emotions such as formant and the cross section of the vocal tract [Muraka 1998; Zhao *et al*. 2001]. As the emotional information of speech signals is more or less related to the meaning of the sentences, the distributing rules and construction characteristics should be attained by analyzing the relationship between emotional speech and neutral speech to avoid the effect caused by the meaning of the sentences.

The global features used in this paper are duration, mean pitch, maximum pitch, average different rate of pitch, average amplitude power, amplitude power dynamic range, average frequency of formant, average different rate of formant, mean slope of the regression line of the peak value of the formant and the average peak value of formant [Zhao *et al*. 2001; Zhao

et al. 2000; Zhao *et al*. 2000]. The duration is the continuous time from start to end in each emotional sentence. It includes the silence, because these parts contribute to the emotion. Duration ratio of emotional speech and neutral speech was used as the characteristic parameters for recognition. The frequency of pitch was obtained by calculating cepstrum. Then the pitch-track was gained, and maximum pitch ($F0_{\max}$), average fundamental frequency ($F0$), average different rate of pitch ($F0_{rate}$) of the envelopes of different emotional speech signals can all be extracted from it. $F0_{rate}$ mentioned here, refers to the mean absolute value of the difference between each frame of speech signal 's fundamental frequencies. The authors used the differences in value of the mean pitch, the maximum pitch and the ratio of $F0_{rate}$ between the emotional and neutral speech as the characteristic parameters. In this paper, the average amplitude power (A) and the dynamic range (A_{range}) are to be taken into account. To avoid the influence of the silent and noisy parts of the speech, the authors only took the mean absolute value of the amplitude into account and all the absolute values must above a threshold. The difference of average amplitude power and the dynamic range between the emotional and neutral speech was used for parameters of recognition. Formant is an important parameter that reflects the characteristics of vocal track. Formant was attained as follows [Zhao *et al*. 2001]. At first, LPC method was applied to calculate 14-order coefficients of linear prediction. Then, the coefficients were used to estimate the track's frequency of the formant by analyzing the frequency average ($F1$), frequency-changing rate ($F1_{rate}$) of the first formant, the average and the average slope of recursive lines of the first four formants. The authors use the difference of $F1$, the last two parameters and the ratio of $F1_{rate}$ between the emotional and neutral speech as the characters in each frame.

The structural features of time series for the emotional sentences used in this paper is maximum value of the pitch in each vowel segment, amplitude power of the corresponding frame, maximum value of the amplitude energy in each vowel segment, pitch of the corresponding frame, duration of each vowel segment and mean value and rate of change of the first three formants. For these parameters, the ratio between the emotional and neutral speech was used as the recognition characters.

In addition to the above features, LPCC, PLP, MFCC are also taken into consideration for precise decision. Figure 2 is the module for feature extraction.

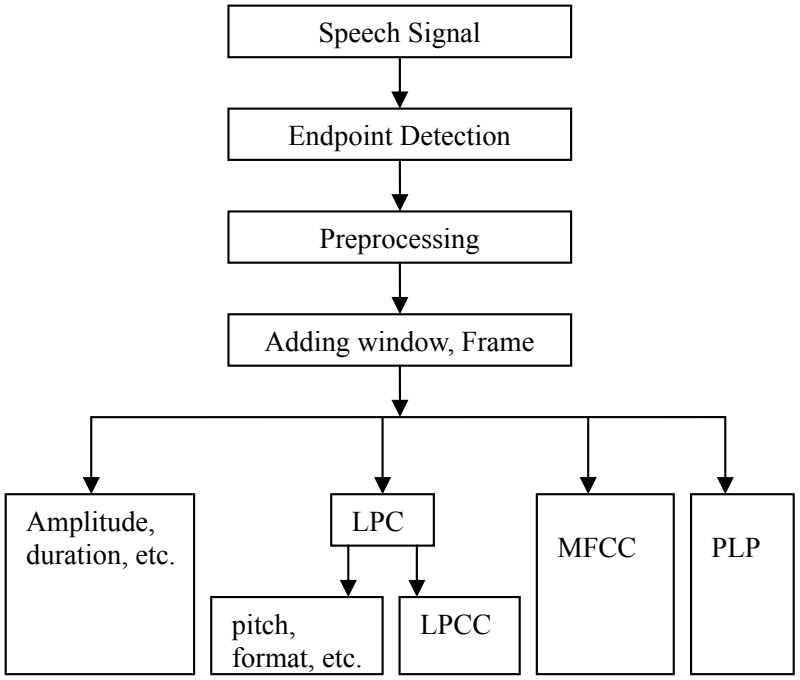


Figure. 2 the module for feature extraction

4. Speech Emotion Recognition based on GMM

GMM can be described as follow:

$$\lambda_i = \{a_i, \mu_i, \Sigma_i\}, \tag{1}$$

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} a_i b_i(\vec{x}), \sum_{i=1}^{M} a_i = 1 \quad , \tag{2}$$

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} \mid \Sigma_i \mid^{1/2}} \cdot \exp\{-\frac{1}{2}(\vec{x} - \mu_i)^t \Sigma_i^{-1}(\vec{x} - \mu_i)\}, \tag{3}$$

where $\vec{x}$ is the D-dimensional feature vector, $b_i(\vec{x})$ $(i = 1,2,...M)$ is the density function of the member $\vec{x}$, $p(\vec{x} \mid \lambda)$ is the probability density function of $\vec{x}$, and a_i satisfies:

$$\sum_{i=1}^{M} a_i = 1 \quad (i = 1,2,...M).$$

The GMM probability function of a speech signal with T frames $X = (\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_T)$ can be denoted as:

$$P(X \mid \lambda) = \prod_{t=1}^{T} p(\vec{x}_t \mid \lambda), \tag{4}$$

or

$$S(X \mid \lambda) = \log P(X \mid \lambda) = \sum_{t=1}^{T} \log p(\vec{x}_t \mid \lambda) . \tag{5}$$

According to the statistical characteristic of likelihood probability (LP) output by Gaussian Mixture Model, the likelihood probability with the best model is generally bigger than that of the other GMM, but due to the existence of variations in speech characteristics and noise, some frames' LP shows a best model that is smaller than that of the others, so the decision may be incorrect. In order to reduce this error recognition rate, some transformation should be introduced to compensate for the likelihood probability, that is, raise the probability with the best model and reduce the probability with the other models. Therefore, a nonlinear compensation transformation is proposed in this paper to solve this problem.

5. Compensation Transformation for GMM

The transformation must satisfy three conditions as follow:

1. The difference of the output probability in different time should be reduced, *i.e.* increase ΔS_1 ;

$$\Delta S_1 = \sum_{\substack{t,k=1 \\ t \neq k}}^{T} \left| \log p(\vec{x}_{tt1} \mid \lambda) - \log p(\vec{x}_k \mid \lambda) \right|$$

2. The difference of the output probability in the same time with different emotion should be increased, *i.e.* increase ΔS_2 ;

$$\Delta S_2 = \sum_{\substack{i,j=1 \\ i \neq j}}^{M} \left| \log p(\vec{x}_t \mid \lambda_i) - \log p(\vec{x}_t \mid \lambda_j) \right|$$

3. The relative value of the output probability should not be changed.

Assuming that $\vec{x}$ is a feature vector, λ_0 is the best model corresponded to $\vec{x}$, and λ_1 is the other model that is mismatched. If the transformation is linear:

$$f[p(\vec{x}_t \mid \lambda_i)] = ap(\vec{x}_t \mid \lambda_i) + b$$

$$f[p(\vec{x} \mid \lambda_0)] - f[p(\vec{x} \mid \lambda_1)] = a[p(\vec{x}_t \mid \lambda_0) - p(\vec{x}_t \mid \lambda_1)] , \tag{6}$$

where $a, b = const$. Here set $a > 0$:

$$p(\vec{x} \mid \lambda_0) \geq p(\vec{x} \mid \lambda_1) \Leftrightarrow f[p(\vec{x} \mid \lambda_0)] \geq f[p(\vec{x} \mid \lambda_1)] , \tag{7}$$

$$p(\vec{x} \mid \lambda_0) \leq p(\vec{x} \mid \lambda_1) \Leftrightarrow f[p(\vec{x} \mid \lambda_0)] \leq f[p(\vec{x} \mid \lambda_1)] . \tag{8}$$

From (7) ~ (8), it is obvious that the linear transformation cannot increase or reduce the LP of the output. The compensation could not be linear transformation, so a nonlinear compensation transformation is proposed; the detailed steps are described as follow:

1. Compute the probability of the t-th feature vector, where N is the number of the emotions, and T is the number of the frames.

$p(\vec{x}_t \mid \lambda_i)$ $(i = 1,2,...N)$, $(t = 1,2,...T)$

2. Normalize $p(\vec{x}_t \mid \lambda_i)$.

$$P(\vec{x}_t \mid \lambda_i) = \frac{p(\vec{x}_t \mid \lambda_i)}{\max p(\vec{x}_t \mid \lambda_i)} \tag{9}$$

3. Compute the output LP.

$$S(\vec{x}_t, \lambda_i) = \frac{[P(\vec{x}_t \mid \lambda_i)]^n}{[P(\vec{x}_t \mid \lambda_i)]^n + b}, \tag{10}$$

where $n = 2 \sim 5$, $b > 1$ and b is always set close to 1.

4. Introduce the compensation: compute the average probability with K former frames.

$$\overline{S}(\vec{x}_{t-K+1,}\vec{x}_{t-K+2,}...,\vec{x}_{t,}\lambda_i) = \frac{1}{K}\sum_{k=1}^{K} S(\vec{x}_{t+k,} \mid \lambda_i) \tag{11}$$

In general, K also has an influence on output probability, here set $K = 2 \sim 5$.

5. Take $\overline{S}(\vec{x}_{t-K+1,}\vec{x}_{t-K+2,}...,\vec{x}_{t,}\lambda_i)$ as the compensation for $S(\vec{x}_t \mid \lambda_i)$.

$$S'(\vec{x}_t \mid \lambda_i) = S(\vec{x}_t \mid \lambda_i) + a_{ti}\delta_{ti}[\overline{S}(\vec{x}_{t-K+1},\vec{x}_{t-K+2},...,\vec{x}_t,\lambda_i) - S(\vec{x}_t \mid \lambda_i)], \tag{12}$$

where $a_{ti} \in [0,1)$, $\delta_{ti} = \begin{cases} 1 & \overline{S}(\vec{x}_{t-K+1},\vec{x}_{t-K+2},...,\vec{x}_t,\lambda_i) > S(\vec{x}_t \mid \lambda_i) \\ -1 & otherwise \end{cases}$.

6. Calculate the joint probability for each model.

$$S(X,\lambda_i) = \sum_{t=1}^{T} \log S'(\vec{x}_t \mid \lambda_i) \tag{13}$$

7. Make the decision of which emotion X belongs to. If $S(X,\lambda_j) = \max_i S(X,\lambda_i)$, then X belongs to λ_j.

Assuming two emotions: λ_0, λ_1 and two vectors: $\vec{x}_1, \vec{x}_2$. Set $T = 2$. The output probability without transformation:

$$S(\vec{x},\lambda_0) = \ln p(\vec{x}_1 \mid \lambda_0) + \ln[p(\vec{x}_2 \mid \lambda_0)], \tag{14}$$

$$S(\vec{x},\lambda_1) = \ln p(\vec{x}_1 \mid \lambda_1) + \ln[p(\vec{x}_2 \mid \lambda_1)]. \tag{15}$$

$$\begin{aligned} \ln P(\vec{x}_1 \mid \lambda_0) + \ln P(\vec{x}_2 \mid \lambda_0) > \ & \ln P(\vec{x}_1 \mid \lambda_1) + \ln P(\vec{x}_2 \mid \lambda_1) \\ \Rightarrow & P(\vec{x}_1 \mid \lambda_0)P(\vec{x}_2 \mid \lambda_0) - P(\vec{x}_1 \mid \lambda_1)P(\vec{x}_2 \mid \lambda_1) > 0 \\ \Rightarrow & P(\vec{x}_1 \mid \lambda_0)^n P(\vec{x}_2 \mid \lambda_0)^n - P(\vec{x}_1 \mid \lambda_1)^n P(\vec{x}_2 \mid \lambda_1)^n > 0, \end{aligned} \tag{16}$$

When $S(\vec{x},\lambda_0) > S(\vec{x},\lambda_1)$, $\vec{x}_i$ (i=1, 2) belongs to λ_0, otherwise belongs to λ_1. The output probability with transformation:

$$S(x_1 \mid \lambda_0) = \log(\frac{P'(\vec{x}_1 \mid \lambda_0)^n}{P'(\vec{x}_1 \mid \lambda_0)^n + b} + \delta_{1,0}\alpha_{1,0}[\frac{P'(\vec{x}_1 \mid \lambda_0)^n}{P'(\vec{x}_1 \mid \lambda_0)^n + b} - \bar{S}(0,\lambda_0)]), \tag{17}$$

$$S(x_2 \mid \lambda_0) = \log(\frac{P'(\vec{x}_2 \mid \lambda_0)^n}{P'(\vec{x}_2 \mid \lambda_0)^n + b} + \delta_{2,0}\alpha_{2,0}[\frac{P'(\vec{x}_2 \mid \lambda_0)^n}{P'(\vec{x}_2 \mid \lambda_0)^n + b} - \bar{S}(1,\lambda_0)]). \tag{18}$$

$S(\vec{x}_1,\lambda_1)$ and $S(\vec{x}_2,\lambda_1)$ are similar to (17)~(18). The decision rule is the same as the one without transformation.

$$\begin{aligned} S(X \mid \lambda_0) - S(X \mid \lambda_1) = & \log(\frac{P'(\vec{x}_1 \mid \lambda_0)^n}{P'(\vec{x}_1 \mid \lambda_0)^n + b} + \delta_{1,0}\alpha_{1,0}[\frac{P'(\vec{x}_1 \mid \lambda_0)^n}{P'(\vec{x}_1 \mid \lambda_0)^n + b} - \bar{S}(0,\lambda_0)]) \\ & + \log(\frac{P'(\vec{x}_2 \mid \lambda_0)^n}{P'(\vec{x}_2 \mid \lambda_0)^n + b} + \delta_{2,0}\alpha_{2,0}[\frac{P'(\vec{x}_2 \mid \lambda_0)^n}{P'(\vec{x}_2 \mid \lambda_0)^n + b} - \bar{S}(1,\lambda_0)]) \\ & - \log(\frac{P'(\vec{x}_1 \mid \lambda_1)^n}{P'(\vec{x}_1 \mid \lambda_1)^n + b} + \delta_{1,1}\alpha_{1,1}[\frac{P'(\vec{x}_1 \mid \lambda_1)^n}{P'(\vec{x}_1 \mid \lambda_1)^n + b} - \bar{S}(0,\lambda_1)]) \\ & - \log(\frac{P'(\vec{x}_2 \mid \lambda_1)^n}{P'(\vec{x}_2 \mid \lambda_1)^n + b} + \delta_{2,1}\alpha_{2,1}[\frac{P'(\vec{x}_2 \mid \lambda_1)^n}{P'(\vec{x}_2 \mid \lambda_1)^n + b} - \bar{S}(1,\lambda_1)]), \end{aligned} \tag{19}$$

Set $a_{10} = a_{20} = a_{11} = a_{22} = const = a$, $p_{ti} = p(\vec{x}_t \mid \lambda_i)$, $\bar{S}_{ti} = \frac{[P(\vec{x}_{t+1} \mid \lambda_i)]^n}{[P(\vec{x}_{t+1} \mid \lambda_i)]^n + b} - \bar{S}(t,\lambda_i)$.

1. $p_{10} = p_{20} = 1$, (16) and (19) can be changed into (20) ~ (21):

$$P_{11}P_{21} < 1 \tag{20}$$

$$\begin{aligned} S(X \mid \lambda_0) - S(X \mid \lambda_1) = & \log[\frac{1}{(1+b)^2} + \frac{a\delta_{1,0}\bar{S}_{00} + a\delta_{2,0}\bar{S}_{10}}{1+b} + a^2\delta_{1,0}\delta_{2,0}\bar{S}_{00}\bar{S}_{10}] \\ & - \log\left[\frac{p_{11}p_{21}}{(p_{11}+b)(p_{21}+b)} + \frac{a\delta_{21}\bar{S}_{11}}{(p_{11}+b)} - \frac{a\delta_{11}\bar{S}_{11}}{(p_{21}+b)} \quad a^2\delta_{11}\delta_{21}\bar{S}_{11}\bar{S}_{21}\right] > 0 \end{aligned} \tag{21}$$

$$\begin{aligned} & \frac{1}{(1+b)^2} - \frac{p_{11}p_{21}}{(p_{11}+b)(p_{21}+b)} + a\left(\frac{\delta_{10}\bar{S}_{00} + a\delta_{20}\bar{S}_{10}}{1+b} - \frac{\delta_{21}\bar{S}_{11}}{(p_{11}+b)} - \frac{\delta_{11}\bar{S}_{11}}{(p_{21}+b)}\right) \\ & \qquad + a^2(\delta_{10}\delta_{20}\bar{S}_{10}\bar{S}_{00} - \delta_{11}\delta_{21}\bar{S}_{11}\bar{S}_{21}) > 0 \end{aligned} \tag{22}$$

s.t.

$$\frac{1}{(1+b)^2} + \frac{a\delta_{10}\bar{S}_{00} + a\delta_{20}\bar{S}_{10}}{1+b} + a^2\delta_{10}\delta_{20}\bar{S}_{10}\bar{S}_{00} > 0$$

$$\frac{p_{11}p_{21}}{(p_{11}+b)(p_{21}+b)}+\frac{a\delta_{21}\bar{S}_{11}}{(p_{11}+b)}-\frac{a\delta_{11}\bar{S}_{11}}{(p_{21}+b)}+a^2\delta_{11}\delta_{21}\bar{S}_{11}\bar{S}_{21}>0$$

where a is small enough to ignore the influence of the second and the third item in (22).

$$\frac{1}{(1+b)^2}-\frac{p_{11}p_{21}}{(p_{11}+b)(p_{21}+b)}=\frac{b^2(1-p_{11}p_{21})+bp_{11}(1-p_{21})+bp_{21}(1-p_{11})}{(1+b)^2(p_{11}+b)(p_{21}+b)}>0 \tag{23}$$

Compared to (20), it can be seen that the LP with transformation is increased.

2. $p_{10}=p_{21}=1$, (16) and (19) can be changed into

$$p_{20}-p_{11}>0 \tag{24}$$

$$S(X\mid\lambda_0)-S(X\mid\lambda_1)=\log\left[\frac{p_{20}}{(1+b)(p_{20}+b)}+\frac{a\delta_{20}\bar{S}_{10}}{1+b}+\frac{a\delta_{10}\bar{S}_{00}p_{20}}{p_{20}+b}+a^2\delta_{10}\delta_{20}\bar{S}_{10}\bar{S}_{00}\right]$$

$$-\log\left[\frac{p_{11}}{(1+b)(p_{11}+b)}+\frac{a\delta_{11}\bar{S}_{01}}{1+b}+\frac{a\delta_{21}\bar{S}_{11}p_{11}}{p_{11}+b}+a^2\delta_{11}\delta_{21}\bar{S}_{11}\bar{S}_{21}]>0 \tag{25}$$

$$\frac{p_{20}}{(1+b)(p_{20}+b)}-\frac{p_{11}}{(1+b)(p_{11}+b)}+a\left(\frac{\delta_{20}\bar{S}_{10}}{1+b}+\frac{a\delta_{10}\bar{S}_{00}p_{20}}{p_{20}+b}-\frac{\delta_{11}\bar{S}_{01}}{1+b}-\frac{\delta_{21}\bar{S}_{11}p_{11}}{p_{11}+b}\right)$$

$$+a^2(\delta_{1,0}\delta_{2,0}\bar{S}_{00}\bar{S}_{10}-\delta_{1,1}\delta_{2,1}\bar{S}_{01}\bar{S}_{11})>0 \tag{26}$$

s.t.

$$\frac{p_{20}}{(1+b)(p_{20}+b)}+\frac{a\delta_{20}\bar{S}_{10}}{1+b}+\frac{a\delta_{10}\bar{S}_{00}p_{20}}{p_{20}+b}+a^2\delta_{10}\delta_{20}\bar{S}_{10}\bar{S}_{00}>0$$

$$\frac{p_{11}}{(1+b)(p_{11}+b)}+\frac{a\delta_{11}\bar{S}_{01}}{1+b}+\frac{a\delta_{21}\bar{S}_{11}p_{11}}{p_{11}+b}+a^2\delta_{11}\delta_{21}\bar{S}_{11}\bar{S}_{21}]>0$$

The first and the second item in (26)

$$\frac{b}{(1+b)(p_{11}+b)(p_{20}+b)}(p_{20}-p_{11}). \tag{27}$$

Compared to (24), (27) has little effect in increasing or reducing probability, except according to the convention: If $P(\bar{x}_1\mid\lambda_0)>P(\bar{x}_2\mid\lambda_0)$, then $P(\bar{x}_1\mid\lambda_1)<P(\bar{x}_2\mid\lambda_1)$. So $\delta_{20}=1,\delta_{21}=-1$, the first and third items in (26) are positive, the second item is far smaller than the first one. Even if the second and the fourth items were negative, the output probability with the best modal would still be bigger than the one with other modals. S_{10} is always bigger than S_{01}, and a is small enough to ignore the fourth item. When the LP of $\vec{x}_1$ with λ_0 and LP of $\vec{x}_2$ with λ_1 is big, the compensation transformation can enlarge the distance between these

two probabilities.

3. $p_{11} = p_{20} = 1$, the analysis is similar to Derivation 2.

6. Experiment Results

In this paper, six people (three male and three female) have taken part in a recording test. They read 27 sentences using five kinds of emotion (happiness, angry, neutral, surprise and sadness), every sentence was read three times, and 2430 sentences were taken as the experiment materials.

GMM with compensation and GMM without compensation are compared first. In the first experiment, globe features and structural features of the time series were utilized. The result is shown in Table 2. In the second experiment, 12 LPCC, 12 MFCC, 16 PLP were utilized. The result is listed in Table 3. Set $K = n = 3, \quad a_{ii} \equiv const = 0.01$

Table 2. the result of the experiments between compensated and uncompensated emotion recognition (globe features and structural features %)

Emotion	Uncompensated GMM	Compensated GMM
Anger	77.6	86.2
Sadness	84.5	99.8
Happiness	73.4	80.0
Surprise	75.8	79.3
Neutral	71.6	77.1

Table 3. the result of the experiments between compensated and uncompensated emotion recognition (LPCC, MFCC, PLP %)

Emotion	Uncompensated GMM	Compensated GMM
Anger	76.3	84.2
Sadness	82.1	97.8
Happiness	79.6	88.3
Surprise	77.8	82.1
Neutral	80.4	87.0

The experiments indicate that the compensation transformation can improve the recognition rate effectively. Angry recognition rate increased 8.2%, sadness recognition rate increased 15.5%, and happiness recognition rate increased 8.5%, surprise recognition rate increased 4%, and neutral recognition rate increased 6%. The selection of K, n, a_{ti} also can improve recognition rate. Here, the authors only selected a set of parameters to explain the effectiveness and robustness of the method. Due to the compensation for GMM, the probability of the output has been stabilized and ΔS_2 has been increased.

Table 4 shows another experiment which compared three methods: KNN, NN [7] and compensated GMM (CGMM).

Table 4. KNN, NN, Compensated GMM (%)

Emotion	KNN	NN	CGMM
Anger	76.0	82.3	86.2
Sadness	82.3	86.0	99.8
Happiness	70.5	71.4	80.0
Surprise	72.2	64.0	79.3
Neutral	78.9	70.6	77.1

Compared to KNN, the recognition rate of anger using CGMM increased 10.2%, sadness increased 17.5%, happiness increased 7.5%, and surprise increased 7.1%, while neutral decreased 1.7%. This decrease doesn't effect the improvement of the whole recognition rate. Compared to NN, the average recognition rate also has been increased about 9.7% using CGMM. The results indicate that CGMM also can improve some other methods to a certain degree.

7. Conclusion and Future Works

In this paper, a method based on GMM with compensation transformation is proposed. In speech emotion recognition, the variations in speech characteristics and noise always influence the recognition results. The common method to solve this problem is conventional preprocessing. As the method in this paper deals with this problem in the recognition stage, the likelihood probability of the output with different models has been increased or decreased to reduce these influences. According to a simple analysis, this compensation transformation can reduce this impact effectively, and the examination results also proved it has better emotion recognition rates. However, the recognition rate of happiness and surprise is still not ideal, and the test materials are too few to further experiments. In further research, the authors will extend the experiment sentences first, then do some studies, such as adding more types of noise and the consideration of gender.

Reference

Austermann, A., N. Esau, L. Kleinjohann, and B. Kleinjohann, "Fuzzy Emotion Recognition in Natural Speech Dialogue," *IEEE International Workshop on Robots and Human Interactive Communication*, 2005, pp. 317-322.

Bhatti1, M. W., Y. Wang, and L. Guan, "A Neural Network Approach for Human Emotion Recognition in Speech," *IEEE Circuits and System, Proceedings of the 2004 International Symposium* ISAS, 2004, vol. 2, pp. 181-184.

Chen, Y.-B., "Automatic Segmentation of Chinese Continuous Speech," In *Proceedings of IEEE Asian Electronics Conference*, 1987, pp. 163-168, Hong Kong, (1987, 09, 1-4).

Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S.Kollias, W. Fellenz, and J.G. Taylor, "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine*, 18(1), 2001, pp. 32-80.

Jiang, D.-N., and L.-H. Cai, "Speech Emotion Classification with the Combination of Statistic Features and Temporal Features," *IEEE International Conference on Multimedia and Expro (ICME)*, June 2004, vol.3, pp. 1967-1970.

Lin, Y. L., and G. Wei, "Speech Emotion Recognition Based on HMM and SVM," In *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, Guangzhou, August 2005, vol. 8, pp.18-21.

Muraka, S.,"Emotional Constituents in Text and Emotional Components in Speech," Ph. D. Theis, *Kyoto: Kyoto Institute of Technology*, Japan, 1998.

Zhao, L., X. Qian, C. Zou, and Z. Wu, "A study on emotional recognition in speech signal," *Journal of Software,* 12(7), 2001, pp. 1050-1055 (in Chinese).

Oppenheim, A. V., C.E. Kopec, and J.M.Tribolet, "Speech Analysis by Homomorphic Prediction," *IEEE Trans*., Vol. ASSP-24, pp. 327-332, 1976.

Ortony, A., and T. J. Turner, "What's Basic About Basic Emotions?" *Psychological Review,* 1990, vol. 97, pp. 315-331.

Park, C.-H., and K.-B. Sim, "Emotion Recognition And Acoustic Analysis From Speech Signal," *IEEE Neural Networks, Proceedings of the International Joint Conference.* vol. 4, 2003 July, pp. 2594-2598.

Rabiner, L. R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Processing of the IEEE,* 1989, 77(2), pp. 257 - 286.

Razak, A.A., R. Komiya, and M. I. Z. Abidin, "Comparison Between Fuzzy and NN Method for Speech Emotion Recognition," *Third International Conference of Information Technology and Applications, 2005, ICITA 2005.* vol. 1, 4-7 July 2005, pp. 297-302.

Shigenaga, M., "Features of Emotionally Uttered Speech Revealed by Discriminant Analysis(VI)," *The preprint of the acoustical society of Japan*, 2-p-18 (1999.9) (in Japan).

Shirasawa, T., and T. Yamamura, "Discriminating Emotion Intended in Speech," *The Preprint of the Acoustical Society of Japan,* HIP: 96-38(1997) (in Japanese).

Zhao, L., X. Qian, C. Zou, and Z. Wu, "A study on emotional feature analysis and recognition in speech signal," *Journal of China Institute of Communications,* 21(10), 2000, pp. 18-25.

Zhao, L., X. Qian, C. Zou, and Z. Wu, "A study on emotional feature extract in speech signal," *Data Collection and Process,* 15(1), 2000, pp. 120-123.

Computational Linguistics and Chinese Language Processing
Vol. 12, No. 1, March 2007, pp. 91-106
© The Association for Computational Linguistics and Chinese Language Processing

The Influence of Reading Styles on Accent Assignment in Mandarin[1]

Mingzhen Bao[*], Min Chu[+], and Yunjia Wang[#]

Abstract

This paper investigates the influences of three different reading styles (*Lyric*, *Critical* and *Explanatory*) to the distribution tendency of sentential accents (classified as rhythmic accent and semantic accent). The comparison among multiple styles is performed in three research domains: high-level constructions, low-level phrases and disyllabic prosodic words. One finds that the assignment of semantic accents shows some differences across reading styles, while the assignment of rhythmic accents does not. Furthermore, the larger the speech unit studied, the stronger the influence is observed, *i.e*. most differences in the assignment of semantic accents are shown in high-level constructions, some are shown in low-level phrases, and none are shown in prosodic words across the three reading styles.

Compared with previous studies, the allocation scheme of semantic accents in the *Explanatory* style is close to that in the neutral style, *i.e*. in high-level constructions, it has a final-accented tendency in theme + rheme (TR), predicate + object(PO) and subject + predicate(SP) constructions, and uniform distribution in adjunct + head constructions. In low-level phrases, the *Explanatory* style exhibits an initial-accented tendency in adjunct + head phrases, but a final-accented tendency in subject + predicate (SP) phrases and predicate + object (PO) phrase. The *Critical* style is adopted to make comments, where semantic focal points are normally on the core subjects and their actions. As a result, more accents are allocated to the subject part in the AS constructions and to the predicate part in the PO constructions. Accordingly, in low-level phrases, more accents go to the heads

---

[1] The work was carried out as an intern in Microsoft Research Asia.

[*] University of Florida, USA
E-mail: joanneb@ufl.edu
[+] Microsoft Research Asia, P. R. China
E-mail: minchu@microsoft.com
[#] Peking University, P. R. China

[Received September 16, 2006; Revised October 25, 2006; Accepted November 1, 2006]

in AN phrases and the predicates in SP phrases. The *Lyric* style helps to express personal emotions in a rhythmic way [Wang 2000]. Such poetry-like rhythm weakens the effect of syntactic constrains, and in many cases, leads to an even distribution of semantic accents in high-level constructions and dense distribution near prosodic boundaries.

Keywords: Reading Style, Sentential Accent, Distribution Tendency, Mandarin

1. Introduction

Stress has been defined as "the degree of force" in terms of speech production [Jones 1976] or as "the degree of loudness" from the viewpoint of speech perception [Trager and Smith 1951]. It has been ranked into different levels of hierarchy, on the top of which is the most salient one, the sentential accent [Zhong and Yang 1999]. In natural speech, sentential accent is distributed to a part of a sentence which is perceived to be more salient than the rest of the sentence. Within the salient part, the sentential accent is assigned to smaller units, first to phrases and words and then to specific syllables. In stress languages, such as English, each word has a primary stress. When the sentence accent is assigned to a polysyllabic word, it is usually obtained by the syllable that holds the primary stress. In tonal languages, such as Mandarin, word stress is usually said to be less salient. According to Chinese phonologists, syllables with four normal tones are all stressed, compared to neutralized syllables. However, from the viewpoint of phonetics, the prominent degree of the "phonologically stressed" syllables varies in polysyllabic words, phrases or sentences. Chao [1979] argued that, in a prosodic unit (a word or a phrase) followed by a pause, the final syllable was primarily accented, the initial one was secondly accented and others in between were weaker than these two. Lin *et al.*'s [1984] experimental study indicated that in most isolated disyllabic words, the final syllables were stressed more heavily than the initial ones.

Sentence accent has been described differently in previous works due to the definition used in the given work. All of these definitions can be classified into two groups if the function of the accent in delivering messages is considered the main factor. Generally speaking, normal accent defined by Newman [1946] and Zhao [1933], or grammar accent defined by Bolinger [1972] and Chomsky [1968], reflecting syntactic or prosodic structures, is predictable with grammatical [Ye 2001] or phonological rules [Luo and Wang 1981]. Contrastive accent, emphatic accent [Lehiste 1970] and logical accent, expressing speaker's special intentions, are hard to be predicted without a deep understanding of the context.

Recently, Chu, Wang, and He studied the accent assignment in Mandarin experimentally. First, they proposed to classify the accents in Mandarin into rhythmic accent (RA) and semantic accent (SA) [Chu *et al*. 2003]. The former serves the function of illustrating the rhythmic structure of an utterance and the later of making the speaker's opinion or intention

prominent. In their works, two experiments were conducted in a speech corpus that contained 300 isolated sentences. In the first experiment, three experts went through the 300 sentences together to identify all accented syllables in the corpus and tagged them as either semantically or rhythmically accented. In order to validate such a classification, they conducted a second experiment. Sixty Mandarin native speakers participated in the experiment. In the results, a relative prominent-level was obtained for each syllable in each sentence. When the results from the two experiments were compared, they found that the syllables tagged as the semantically accented had significantly higher prominent-level than those tagged as rhythmically accented. Both types of accented syllables had much higher prominent-level than the unaccented syllables. Furthermore, some syllables judged as to have both the semantic and the rhythmic accents in the first experiment achieved the highest prominent-level in the second experiment. All these results supported the separation of semantic accent from rhythmic accent. In the follow-up studies [Wang *et al.* 2003a, b], they found that the rhythmic accent tended to be assigned to the final syllable within a prosodic word and a prosodic phrase , while no patterns were found for the distribution of semantic accent. Later, in a study of semantic accents alone, Wang *et al.* [2003c] found that the distribution tendency of semantic accent changed with the speech unit studied. For example, in a low-level phrase or a prosodic word, semantic accent was often found in the modifiers when it had a modifier-head structure. However, such a tendency did not show up in high-level constructions.

Conclusions in [Chu *et al*. 2003; Wang *et al*. 2003a, b, c] were drawn from the observation of independent sentences read with a neutral intonation. In this paper, the authors extend the study into affective speech. The accent assignment tendency is compared among three reading styles to find out whether reading styles have any influence on accent assignment in the research domains of prosodic words, low-level phrases and high-level constructions.

2. Data Preparing and Processing

2.1 The Speech Corpus

Seven articles were selected for this study, in which, two were lyric essays by famous Chinese writers, two were remarks on a newly-published novel and a newly-drawn policy, and three were objective illustrations about the weather, the stock market and a new law, respectively. These articles were read by the same voice talent who also read the independent sentences studied in Wang, Chu, and He's works. Unlike previous recording sessions where the voice talent was asked to read sentences with a neutral intonation, this time, she was requested to choose a proper reading style for each type of article according to her understanding of these articles. According to the voice talent, she used different reading styles for the three groups of

articles. These styles could be discriminated by listeners in an informal listening test though they could not give a clear linguistic term for each type. In this paper, the authors name the three reading styles as *Lyric*, *Critical* and *Explanatory*, respectively. The difference in speech rate shown in Table 1 is an acoustic support for the division of the reading styles [Fackrell *et al*. 2000]. The *Lyric* style was presented the slowest, the *Critical* style the fastest, and the *Explanatory* style in the middle.

Table 1. Comparison of the speech rates of the three reading styles.

Reading Style	Lyric	Criti.	Exp.
Total num. of syllables	897	697	1450
Speech Rate (char per minute)	210	250	230

2.2 Annotation of Accents

The locations and types of accents within the seven articles were annotated by two graduate students majoring in linguistics, who were interested in phonetics. After listening to the recordings, they were asked to identify all accents in the speech corpus and assign a type (rhythmic, semantic or both) to each with the same guidelines (listed in Table 2) that were used in Wang, Chu, and He's studies on accent assignment in neutral sentences [Chu *et al*. 2003; Wang *et al*. 2003a, b, c].

Table 2. Guidelines for identifying accents and their types

1.	Annotators can listen to a sentence as many times as they want;
2.	At least one accent should be labeled in each sentence, and it can be semantically accented, rhythmically accented or both;
3.	Multiple accents are allowed in one utterance and there is no hard threshold for the maximum number per utterance.

Before the formal annotation, the two annotators were trained with a subset of materials annotated in previous studies [Chu *et al*. 2003; Wang *et al*. 2003a, b, c]. The training took two steps, annotation and discussion to improve the across-person agreement. First, they annotated accents independently according to the definitions and guidelines given in [Chu *et al*. 2003]. The initial agreement-ratio on both the location and the type of accents was only 56.4%. Then, they discussed all of the differences and got access to the annotation obtained in the previous works. After the discussion, they achieved agreement on most of the different cases. Finally, they labeled another subset of the isolated sentences independently. This time, the agreement-ratio increased to 67.6%. Such a training cycle was repeated three times. The first training session brought about an 11% increase in agreement. However, the second and the third sessions did not bring much improvement. The highest agreement-ratio achieved was about 70%. Since the agreement-ratio was not as high as expected, the authors will keep the

discussion part for the annotation of the new corpus. The two annotators labeled accents in the seven articles independently first and discussed cases where different opinions appeared. For a few cases where they could not agree with each other, a third person was invited to make the final decision. As a result, each syllable in the seven articles obtained one of the four accent labels, UA — unaccented, SA — semantic accented, RA — rhythmic accented, SRA — semantic and rhythmic accented.

2.3 Annotation of Syntactic Structure

In Chinese, many syntactic structures in the sentences level can be used recursively to construct phrases and words. In previous works, it was found that the accent assignment has different tendencies in different levels of constituents in the neutral reading style. In this work, the accent assignment under different reading styles is studied in the same three levels, including the high-level construction, the low-level phrase and the prosodic word. The anchors for the three constituents are the top chunks, the prosodic words, and the syllables in a sentence.

2.3.1 High Level Construction

The largest speech unit the authors are interested in is the high-level construction, and the anchor for identifying a high-level phrase is the top chunk of the sentence. First, a sentence is chunked into several linearly succeeding components, including ① sentence adjunct；②subject adjunct; ③ subject; ④predicate adjunct; ⑤predicate; ⑥object adjunct; and ⑦object. Then, immediate constructions that are formed by these chunks are identified and labeled as one of the following structures: ①TR — theme + rheme; ②PO — predicate + object; ③SP — subject + predicate; ④AO — adjunct + object; ⑤AS — adjunct + subject; ⑥AP — adjunct + predicate. The authors will discuss which parts of certain types of construction tend to be accented in different reading styles. An example of the top chunk level annotation is shown in Figure 1 (a).

2.3.2 Low Level Phrase

The second speech unit investigated was the low-level phrase, the anchor of which is prosodic words in a sentence. The authors wanted to find out whether the rules for accent assignments in low-level phrases are the same as those used in high-level constructions, and whether speaking styles exert the same effect on them. First, the authors scanned each prosodic word in a sentence from left to right and identified the immediate carrying phrase of the target word. Then, the structure of this phrase was analyzed, and the target word was labeled with its role in the carrying phrase. Seven types of structures were annotated for the low-level phrases, which are ①SP — subject + predicate；②AN — attribute + noun head; ③ PC — predicate +

complement; ④AV — adverbial + verb head; ⑤PO — predicate + object; ⑥CO — coordinative construction; ⑦ PP — preposition phrase. The role of each word in its immediate carrying phrase was labeled as the "attribute in AN" or the "verb in AV", etc. An example is given in Figure 1 (b) where the prosodic word "高处" is an "attribute in AN" in its carrying phrase "高处丛生的灌木" and the prosodic word "丛生的" is the "attribute in AN" in its carrying phrase "丛生的灌木". By comparing the frequency of how often the "attribute in AN" receives accents with that of the "head in AN", one can figure out the accent assignment tendency in AN phrases.

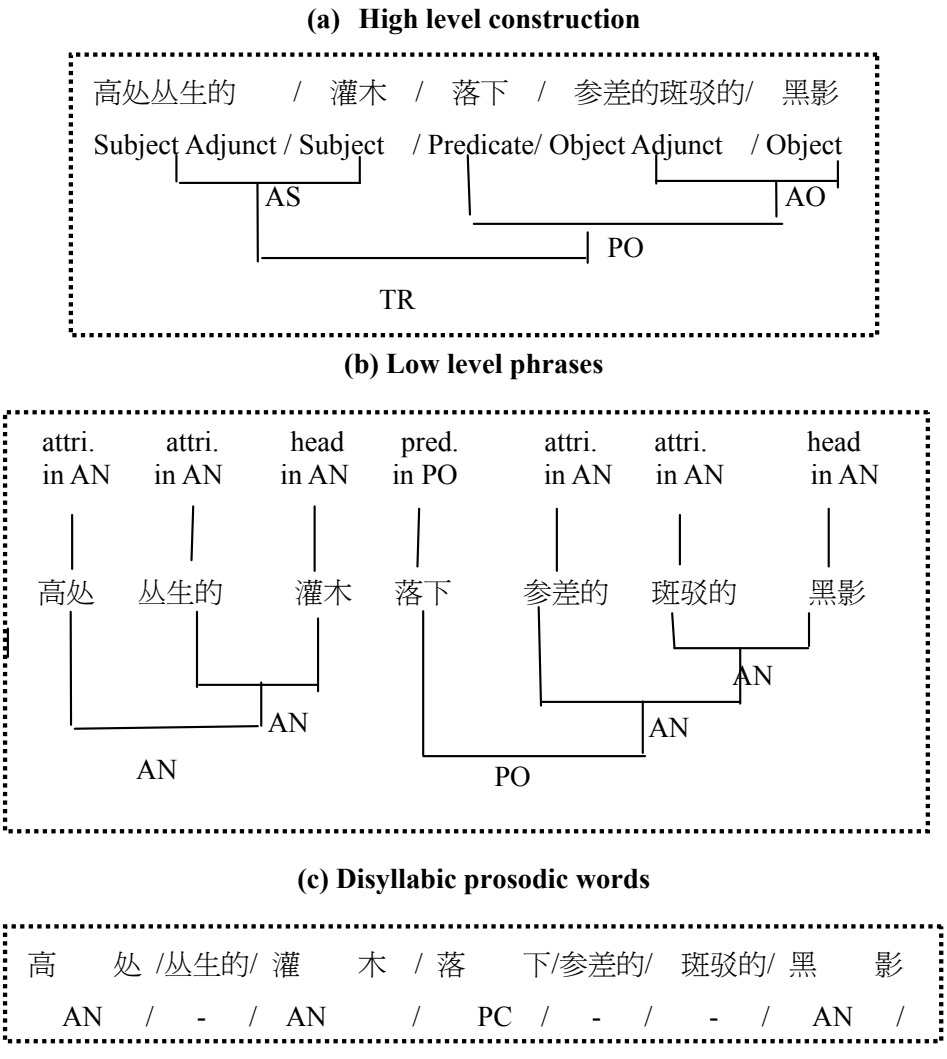


Figure 1. An example of structural labeling in the sentence "高处丛生的灌木落下参差的斑驳的黑影"("Tufty shrubs in the upland cast spotted irregular shadows.")。

2.3.3 Prosodic Word

The third level of speech unit studied was the prosodic word. In this paper, the authors only focus on disyllabic words since they are the most common Chinese words. The same seven types of syntactic structure used in Section 2.3.2 are annotated for words and an example is shown in Figure 1 (c).

In the next section, the authors compare accent assignment tendency in the three level units among the three reading styles respectively.

2.4 Indicators for Accent Assignment

Since no limitation has been put on the total number of accents per sentence, it is possible that more than one word in a top chunk is accented. Therefore, the comparison between the numbers of accented words in the two immediate chunks of a high-level construction does not tell the accent tendency directly (*i.e.* which part of the construction tends to receive accents). Similarly, in low-level phrases, the total number of words in one type of constituents is often different from that of the other type of constituents. For example, in Figure 1(b), there are 4 "attri. in AN" while only two "head in AN" in low level phrases. Thus, the ratio of the number of accented words in "attri. in AN" class to that in the "head in AN" class does not directly reflect the accent tendency either (*e.g.*, the ratio is 2:1, if accents are distributed normally among all words).

To describe the accent tendency in a better way, an *accent indicator* (*AI*) is defined as the ratio of the number of obtained accents to the expected number of accents in a certain class of words as in (1). It shows the possibility for a class of chunks or words to obtain sentential accents.

$$AI=N_r / N_p \tag{1}$$

N_r is the number of accents obtained by a class of words and N_p is the expected number of accents for the class under the assumption that all accents are distributed normally among all syllables in the corpus. N_p is calculated by (2) and (3).

$$N_p = N_w \times P \tag{2}$$

$$P= N_s / N_a \tag{3}$$

N_a is the number of syllables in the corpus, while N_s is the number of accented syllables in it. P indicates the possibility of a syllable to obtain a sentential accent under the assumption of normal distribution. N_w is the number of syllables in a class studied.

$AI>1$ means that the possibility for the corresponding class to obtain accents is above the average, *i.e.* it tends to obtain more sentential accents. $AI < 1$ means the opposite and $AI = 1$

means it has the average possibility of being accented.

To illustrate the accent tendency within a certain construction, *i.e*. to answer the question of which part between the two immediate constituents of a construction is more often to be accented, an *accent indicator ratio* (*AIR*) is defined as the ratio of the *AI* of the initial component to that of the final. *AIR*>1 describes an initial-accented tendency, while *AIR*<1 argues for a final-accent tendency. *AIR*=1 means the two components within the construction have equal chance to be accented.

Since the initial parts of disyllabic words always share the same number of syllables as the final parts, *AI* is not needed in studying disyllabic words. *AIR* is defined as the ratio of the number of accented syllables in the two parts for a given word category.

3. Results and Analyses

The *AI* and *AIR* described in subsection 2.4 are calculated for the three prosodic units and the results for the three reading styles are compared in subsection 3.1, 3.2 and 3.3, respectively.

3.1 Accent Assignment in High Level Constructions

AIs of semantic accents (SA) in the six types of sentence constructions are calculated for the three reading styles in Table 3(a). The corresponding *AIRs* in each type of construction are given in Table 3(b). From the two tables, a weak final-accented tendency (*AIR<1*) is observed in TR, PO and SP constructions in most reading styles, *i.e*., when semantic accents are assigned to these constructions, it often goes to the rhemes, objects, or predicates. This observation tallies with the previous findings in neutral speech. Some exceptions lie in the TR, SP constructions under the *Lyric* style and the PO construction under the *Critical* style, where semantic accents are uniformly distributed.

When looking into the constructions with adjunct + head structures (AO, AS and AP), one finds that semantic accents have different tendencies under different reading styles. For example, adjuncts in noun-head phrases (AO and AS) tend to be accented in the *Critical* style, while, in the *Lyric* style, all heads tend to be accented. The *Explanatory* style shows no strong tendency in both AO and AP phrases and has initial-accented tendency in AS phrases. Compared with the results in the previous study of neutral speech, The *Explanatory* style shows most similarity tendency to the neutral style.

AIs and *AIRs* of rhythmic accents (RA) in the six constructions are calculated for the three reading styles in Table 3(c) and (d). This shows that the distribution tendencies of rhythmic accents are quite similar across the three reading styles in most construction types, *i.e*. they are evenly distributed in TR and SP constructions and final-accented in AO, AS and AP constructions. It is worth noticing that most *AIR*s in AO and AS constructions are smaller

than those in AP, which indicates chunks within AO and AS constructions are more likely to be tightened up into one prosodic unit than those of AP. This conclusion is consistent with the one drawn in [Chu *et al*. 2003]. The main exceptions in rhythmic accent assignment are in PO constructions. There is a tendency towards final-accented constructions appearing in the *Lyric* and the *Critical* styles, yet a tendency towards initial-accented constructions in the *Explanatory* style.

Table 3. Accent indicators in six types of constructions under three reading styles

(a) Accent indicators for semantic accents

Construction type	Chunk property	Reading styles		
		Lyric	Criti.	Exp.
TR	Theme	1.00	0.80	0.90
	Rheme	1.00	1.10	1.10
PO	Predicate	0.50	1.00	0.60
	Object	1.30	1.10	1.20
SP	Subject	1.10	0.70	0.80
	Predicate	1.20	1.00	1.10
AO	Adjunct	1.05	1.20	1.23
	Object	1.35	1.06	1.17
AS	Adjunct	0.81	1.13	0.69
	Subject	1.07	0.47	0.97
AP	Adjunct	0.65	1.07	1.00
	Predicate	0.64	0.92	0.99

(b) Accent indicator ratios for semantic accents

Construction type	Reading styles		
	Lyric	Criti.	Exp.
TR	1.00	0.73	0.82
PO	0.38	0.92	0.50
SP	0.92	0.70	0.73
AO	0.78	1.12	1.04
AS	0.75	2.41	0.72
AP	1.02	1.16	1.01

(c) Accent indicators for rhythmic accents

Construction type	Chunk property	Reading styles		
		Lyric	Criti.	Exp.
TR	Theme	1.00	1.00	1.00
	Rheme	1.00	1.00	1.00
PO	Predicate	0.70	0.70	1.20
	Object	1.20	1.30	1.00
SP	Subject	1.20	0.60	0.70
	Predicate	1.10	0.70	0.70
AO	Adjunct	0.35	0.79	0.33
	Object	1.41	1.91	1.50
AS	Adjunct	0.00	0.32	0.42
	Subject	1.33	1.61	1.18
AP	Adjunct	0.47	0.67	0.28
	Predicate	0.82	0.75	1.22

(d) Accent indicator ratios for rhythmic accents

Construction type	Reading styles		
	Lyric	Criti.	Exp.
TR	1.00	1.00	1.00
PO	0.58	0.54	1.20
SP	1.10	0.86	1.00
AO	0.25	0.41	0.22
AS	0.00	0.20	0.36
AP	0.57	0.90	0.23

3.2 Accent Assignment in Low Level Phrases

Since CO, PP and PC phrases appeared only a few times in each reading style, only four types of phrases, *i.e*. AN, AV, PO and SP, are studied in this paper. *AI* and *AIR* in the four categories are calculated separately under the three reading styles. The results are listed in Table 4, in which, (a) and (b) are *AI* and *AIR* for semantic accent, and (c) and (d) are for rhythmic accent.

Table 4. Accent indicators in four types of low level phrases under three reading styles

(a) Accent indicators for semantic accents

Phrase type	Word property	Reading styles		
		Lyric	Criti.	Exp.
AN	Attribute	1.22	1.50	1.39
	Head	1.52	0.95	0.98
AV	Adverbial	0.81	1.05	1.17
	Head	1.05	1.09	0.93
PO	Predicate	0.65	0.34	0.44
	Object	1.61	1.54	0.57
SP	Subject	0.74	0.33	0.63
	Predicate	0.78	1.96	2.04

(b) Accent indicator ratios for semantic accents

Phrase type	Reading styles		
	Lyric	Criti.	Exp.
AN	0.81	1.58	1.42
AV	0.76	0.96	1.26
PO	0.40	0.22	0.77
SP	0.95	0.17	0.31

(c) Accent indicators for rhythmic accents

Phrase type	Word property	Reading styles		
		Lyric	Criti.	Exp.
AN	Attribute	0.27	0.18	0.10
	Head	2.24	2.35	2.10
AV	Adverbial	0.42	0.25	0.19
	Head	1.66	1.29	1.57
PO	Predicate	0.22	0.11	0.42
	Object	2.02	2.84	2.42
SP	Subject	0.94	1.71	0.95
	Predicate	2.34	3.85	2.42

(d) Accent indicator ratios for rhythmic accents

Phrase type	Reading styles		
	Lyric	Criti.	Exp.
AN	0.12	0.08	0.05
AV	0.25	0.18	0.12
PO	0.11	0.04	0.17
SP	0.40	0.44	0.39

From Table 4(a)-(b), the results among reading styles show more diversity.

(a) In AN phrases, all *AIRs* except that under the *Lyric* style, are larger than 1. This shows the semantic accent tends to be assigned to the adjunct under the *Critical* and the *Explanatory* styles and to the head under the *Lyric* style.

(b) In AV phrases, *AIRs* in the *Explanatory* style show a tendency toward being initial-accented, while the *Lyric* style has a tendency toward being final-accented. The chances of being accented for both components under the *Critical* style are almost the same.

(c) In PO phrases, all *AIRs* are smaller than 1, *i.e.*, PO phrases have final-accented tendency. Among the three reading styles, the final-accented tendency is weakest under the *Explanatory* style.

(d) Under the *Critical* and the *Explanatory* styles, SP phrases show strong final-accented tendency. Yet, under the *Lyric* style, the two immediate components of SP phrases have an equal chance of obtaining semantic accents.

Comparing these results with those in previous studies, one can see that both the *Critical* and the *Explanatory* styles show the same initial-accented tendency in AN phrases as the neutral style, while the initial-accented tendency in AV phrases is weakened in the *Critical* style. The *Lyric* style has the opposite tendency in both AN and AV phrases. The two immediate components of PO phrases have an equal chance of obtaining semantic accents in the neutral style. However, both the *Critical* and the *Lyric* styles have rather strong final-accented tendency in PO phrases, and such a final-accented tendency is weakened in the *Explanatory* style. The *Lyric* style shows similar distribution of sentential accents in SP phrases to the neutral style, but the other two styles have strong final-accented tendency.

For rhythmic accent, a final-accented tendency is observed unanimously in Table 4 (d), regardless of reading styles. This is consistent with the conclusions drawn from independent neutral sentences [Chu *et al*. 2003] [Wang *et al*. 2003b], and it further demonstrates that the final-accented tendency of rhythmic accent is not influenced by reading styles. An interesting phenomenon is presented in that SP phrases in all reading styles always have the largest *AIRs* among all types of phrases, *i.e*., the final-accented tendency is comparatively weak in SP

phrases. A possible reason is that, when words are grouped into prosodic phrases, the relationship between the subjects and the predicates in SP phrases is not as close as in other phrases so the two components are often grouped into different prosodic phrases [Wang *et al*. 2003c].

3.3 Accent Assignment in Disyllabic Prosodic Words

Since initial parts of disyllabic words share the same number of syllables as final parts, no *AI* is adopted. *AIRs* are calculated for word types with more than 10 observations in the speech corpus. The results are listed in Table 5, in which, (a) is for semantic accent and (b) is for rhythmic accent.

Table 5. Accent indicator in three types of prosodic words under three reading styles

(a) Accent indicator ratios for semantic accents[2]

Phrase type	Reading styles		
	Lyric	Criti.	Exp.
AN	2.91	4.75	8.57
AV	+∞	1.33	2.63
PO			1.86

(b) Accent indicator ratios for rhythmic accents[3]

Word type	Reading styles		
	Lyric	Criti.	Exp.
CO	0.24	0.06	0.22
AN	0.16	0.21	0.07
PO			0.43
AV			0.06

From Table 5(a)-(b), the initial-accented tendency for semantic accent (*AIR*s>1) and the final-accented tendency for rhythmic accent (*AIR*s<1) are consistently observed in the three reading styles. These observations comply with the previous study on accent distribution in the neutral style. Therefore, one can conclude that accent distribution within prosodic words is seldom affected by reading styles.

[2] "+∞" means stress is always distributed to initial syllables without an exception.

[3] Blank cells in Table 5 indicate no enough observations are available for certain cases.

4. Conclusions and Discussions

This paper investigates the influence of reading styles on the accent assignment within high-level constructions, low-level phrases and prosodic words. The results show that (1) semantic accents are more affected by reading styles than rhythmic accents and (2) more significant influences are observed in larger speech units (such as the high-level constructions and the low-level phrases) than in smaller units (such as prosodic words). In detail, 1) Semantic accents show a strong initial-accented tendency in all types of prosodic words across different reading style, while, rhythmic accents unanimously demonstrate a final-accented tendency in prosodic words; 2) In high-level constructions, semantic accents tend to be allocated to the final constituents within TR, PO, SP and AS structures in the *Explanatory* style; within TR and SP structures in the *Critical* style and PO, AO and AS in the *Lyric* style, and they are allocated to the initial constituents within the AO, AS and AP structure in the *Critical* style. Compared with previous study in neutral speech, the *Explanatory* style has similar impact on accent allocation in high-level constructions to the neutral style. The *Critical* style weakens the final-accented tendency in PO constructions and demonstrates strong initial-accented tendency in AS constructions. The *Lyric* style presents more diversity with no significant tendency in TR, SP and AP constructions, and initial-accented tendency in PO, AO and AS constructions; 3) In low-level phrases, semantic accents are often allocated to the final parts within PO and SP phrases. Yet, such a final-accented tendency is weaker for the *Lyric* style in SP phrases and the *Explanatory* style in PO phrases. In AN phrases, the noun-heads are often accented in the *Explanatory* and the *Critical* styles, yet, accents normally go to the adjuncts in the *Lyric* style. Both the *Lyric* and the *Critical* styles demonstrate a final-accented tendency in AV phrases where an initial-accented tendency is observed in the *Explanatory* style.

These results are consistent with the theory of ornate form [Milic 1965]: to deliver the attitude of a speaker through speaking styles. Listeners and speakers share an accent system as a convention in which listeners know to go to accented items to find information which the speaker is particularly attentive to produce. Therefore, semantic accent is more closely related to reading styles and easier to be influenced.

In the *Explanatory* style, the speaker's task is to present messages clearly and concisely with an objective tone. This is also a regular way to deliver independent neutral sentences where syntactic constraints work actively. Therefore, the overall tendency for semantic accent assignment in this style is rather close to that in neutral style and is mainly constrained by the syntactic and the prosodic structures of a sentence.

The *Critical* style is adopted to make comments, where semantic focuses are normally on the core subjects and their actions. As a result, more accents are allocated to the subject part in

the AS constructions and to the predicate part in the PO constructions. Accordingly, in low-level phrases, more accents go to the heads in AN phrases and the predicates in the SP phrase. However, in AV phrases, both the adjuncts and the verbs have equal chance to be accented. A possible reason for this is that the manners for actions to take place sometimes also play an important role in the discourse. The authors do not have a good explanation for why accents tend to be allocated to the objects in PO phrases.

The *Lyric* style helps to express personal emotions in a rhythmic way [Wang 2000]. Such poetry-like rhythm weakens the effect of syntactic constrains and, in many cases, leads to an even distribution of semantic accents in high-level constructions. For low-level phrases, more semantic accents are observed near prosodic boundaries to meet the requirement of rhyme-scheme, and accordingly final-accented tendencies are presented in AN and AV phrases.

References

Bolinger, D., "Accent is Predictable (if You're a Mind-Reader)," *Language*, 48(3), 1972, pp. 633-644.

Chao, Y. R., *A Grammar of Spoken Chinese*, being translated by S. Lu, Commercial Press (Beijing), 1979. (In Chinese).

Chomsky, N., and H. Morris, *The Sound Pattern of English*, New York: Harper and Row, 1968.

Chu, M., Y.J. Wang, and M.Z. Bao, "Local Grammatical Constraints and Length Constraints for Forming Base Prosodic Phrase in Mandarin," In *Proceedings of the 6th National Conference of Modern Phonetics*, 2003, Tianjin, P.R. China, pp. 161. (in Chinese).

Chu, M., Y.J. Wang, and L. He, "Labeling Stress in Continuous Mandarin Speech Perceptually," In *Proceedings of the 15th International Congress of Phonetic Sciences*, 2003, Barcelona, Spain, pp. 2095-2098.

Jones, D., *An Outline of English Phonetics*, Cambridge University Press, 1976.

Fackrell, J., H. Vereecken, J.-P. Martens, and B. Van Coile, "Prosodic Variation with Text Type," *IEE Seminar on State of the Art in Speech Synthesis*, 2000, pp. 5/1-5/9.

Lehiste, I., *Suprasegmentals*, M.I.T.Press, 1970.

Lin, M. C., J. Z. Yan, and G. H. Sun, "A Primary Experiment on the Stress Pattern of Normal Disyllabic Words in Mandarin," *Dialect*, 1, 1984, pp. 57-73. (In Chinese).

Milic, L, T., "Theories of Style and Their implications for the Teaching of Composition," *College Composition and Communication,* 16, 1965, pp. 66-69, 126.

Luo, C.P., and J. Wang, *An Outline of General Phonetics*, Commercial Press, 1981.

Newman, S., "On the Stress System of English," *Word*, 2, 1946, pp. 171-87.

Trager, G.L., and H.L.J. Smith, *An Outline of English Structure*, Norman, Oklahoma: Battenburg Press, 1951.

Wang, Y.J., M. Chu, and L. He, "Classification and Distribution of Sentence Stress in Mandarin," *Journal of Psychology*, 35(6), 2003a, pp. 734-742. (in Chinese).

Wang, Y.J., M. Chu, and L. He, "An Experimental Study on the Distribution of Focus Accent in Mandarin," *Chinese Teaching in the World,* 2, 2003b, pp. 86-98. (in Chinese).

Wang, Y.J., M. Chu, and L. He, "Location of Sentence Stresses within Disyllabic Words in Mandarin," In *Proceedings of the 15th International Congress of Phonetic Sciences,* 2003c, Barcelona, Spain, pp. 1827-1830.

Wang, Z.S., "Development and Changes in Modern Essays," *Chinese Literary Research,* 4, 2000, pp. 11-20. (in Chinese).

Ye, J., *Grammatical Functions of Chinese Prosody*, East China Normal University Press, 2001. (in Chinese).

Zhao, Y.R., "Tone and Intonation in Chinese," *Bulletin of the Institute of History and Philology*, 4, 1933, pp. 121-134.

Zhong, X.B., and Y.F. Yang, "Foreign Researches on Prosodic Features and Stresses," *Acta Psychologica Sinica,* 4, 1999, pp. 468-475. (in Chinese).