Discriminating Neutral and Emotional Speech using Neural Networks

Sudarsana Reddy Kadiri[1], **P. Gangamohan** [2] **and B. Yegnanarayana** [3]
Speech and Vision Laboratory,
Language Technologies Research Center,
International Institute of Information Technology-Hyderabad, India.
[1]sudarsanareddy.kadiri@research.iiit.ac.in, [2]gangamohan.p@students.iiit.ac.in,
[3]yegna@iiit.ac.in

Abstract

In this paper, we address the issue of speaker-specific emotion detection (neutral vs emotion) from speech signals with models for neutral speech as reference. As emotional speech is produced by the human speech production mechanism, the emotion information is expected to lie in the features of both excitation source and the vocal tract system. Linear Prediction residual is used as the excitation source component and Linear Prediction Coefficients as the vocal tract system component. A pitch synchronous analysis is performed. Separate Autoassociative Neural Network models are developed to capture the information specific to neutral speech, from the excitation and the vocal tract system components. Experimental results show that the excitation source carries more information than the vocal tract system. The accuracy neutral vs emotion classification using excitation source information is 91%, which is 8% higher than the accuracy obtained using vocal tract system information. The Berlin EMO-DB database is used in this study. It is observed that, the proposed emotion detection system provides an improvement of approximately 10% using excitation source features and 3% using vocal tract system features over the recently proposed emotion detection which uses the energy and pitch contour modeling with functional data analysis.

Keywords: Excitation Source, Vocal Tract System, Linear Prediction (LP) Analysis, Autoassociative Neural Network.

1 Introduction

Speech is produced by the human speech production mechanism, and it carries the signature of the speaker, message, language, dialect, age, gender, context, culture, and state of the speaker such as emotions or expressive states. Extraction of these elements of information from the speech signal depends on identification and extraction of relevant acoustic parameters. Information present in the speech signal, including emotional state of a speaker, has its impact on the performance of speech systems (Athanaselis et al., 2005).

In this study, emotion detection refers to, identification of whether the speech is neutral or emotional. Emotion recognition refers to determining the category of emotion, i.e., anger, happy, sad, etc. The focus in this study is on detection of presence of emotional state of a speaker with the use of reference models for neutral speech. Motivated by a broad range of commercial applications, automatic emotion recognition from speech has gained increasing research attention over the past few years. Some of the applications for emotion recognition system are in the fields of health care, call centre services and also for developing speech systems such as automatic speech recognizer (ASR) to improve the performance of dialogue systems (Athanaselis et al., 2005; Mehu and Scherer, 2012; Cowie et al., 2001; Morrison et al., 2007).

Extraction of features from speech signal that characterize the emotion content of speech, and at the same time do not depend on the lexical content is an important issue in emotion recognition (Schuller et al., 2010; Luengo et al., 2010; Scherer, 2003; Williams and Stevens, 1972; Murray and Arnott, 1993; Lee and Narayanan, 2005). From (Schuller et al., 2010; Hassan and Damper, 2012; Schuller et al., 2013; Schuller et al., 2011), it is observed that there is no clear understanding

D S Sharma, R Sangal and J D Pawar. Proc. of the 11th Intl. Conference on Natural Language Processing, pages 214–221,
Goa, India. December 2014.

on what type of features can be used for emotion recognition task. Brute force approach involves extracting as many features as possible, and use these in the experiments, sometimes using feature selection mechanisms to choose appropriate subset of features (Schuller et al., 2013; Schuller et al., 2011; Schuller et al., 2009; Zeng et al., 2009). These features can be broadly classified as prosodic features (pitch, intensity, duration), voice quality features (jitter, shimmer, harmonic to noise ratio (HNR)), spectral features (Mel Frequency Cepstral Coefficients (MFCCs), Linear Prediction Cepstral Coefficients (LPCCs)), and their statistics such as mean, variance, minimum, maximum, range (Zeng et al., 2009; Schuller et al., 2011; Schuller et al., 2009; **?**; Eyben et al., 2012). A limitation of this approach is the assumption that every segment in the utterance is equally important. Studies have shown that emotional information is not uniformly distributed in time (Jeon et al., 2011; Lee et al., 2011; Shami and Verhelst, 2007).

In (Busso et al., 2009; Bulut and Narayanan, 2008; Arias et al., 2014; Arias et al., 2013; Busso et al., 2007), authors observed that a robust neutral speech models can be useful in contrasting different emotions expressed in speech. Emotion detection study was made by creating acoustic spectral features of neutral speech with HMMs (Busso et al., 2007). In (Busso et al., 2009), authors used the pitch features of neutral speech to discriminate the emotions using the Kullback-Leibler distance. It was observed that gross pitch contour statistics such as mean, minimum, maximum and range are prominent than pitch shape. Recently, emotion detection is performed using functional data analysis (FDA) (Arias et al., 2014; Arias et al., 2013). In this approach, pitch and energy contours of neutral speech utterance are modeled using FDA. In testing, pitch and energy contours are projected onto the reference bases, and their projections are used to discriminate neutral and emotional speech. Similar studies were made to model the shape of pitch contour of emotional speech by analyzing the rising and falling movements (Astrid and Sendlmeier, 2010). One limitation with the studies (Arias et al., 2014; Arias et al., 2013) is that, all the utterances should be temporally aligned with the Dynamic Time Warping and it may not be realistic for most of the situations.

Here, we propose an approach based on AANN (Yegnanarayana and Kishore, 2002) to detect whether a given utterance is neutral or emotional speech. The detection of emotional segments or emotion events may help the current approaches in automatic emotion recognition. This approach avoids the interrelations among the lexical content used, language and emotional state across varying acoustic features. The discrimination capabilities of AANN models are exploited in various areas of speech such as speaker identification, speaker verification, speaker recognition, language identification, throat microphone processing, audio clip classification etc (Reddy et al., 2010; Murty and Yegnanarayana, 2006; Yegnanarayana et al., 2001; Mary and Yegnanarayana, 2008; Bajpai and Yegnanarayana, 2008; Shahina and Yegnanarayana, 2007).

This present work is based on our previous work (Gangamohan et al., 2013) for capturing the deviations of emotional speech from neutral speech. In that paper (Gangamohan et al., 2013), it was shown that the excitation source features extracted in the high signal to noise ratio (SNR) regions of the speech signal (around the glottal closure) capture the deviations of emotional speech from neutral speech. This paper presents a framework to characterize the high SNR regions of the speech signal using the knowledge of speech production mechanism. In (Reddy et al., 2010; Murty and Yegnanarayana, 2006; B. Yegnanarayana and S. R. Mahadeva Prasanna and K. Sreenivasa Rao, 2002), the authors showed the importance of processing the high SNR regions of speech signal for various applications such as speaker recognition (Reddy et al., 2010; Murty and Yegnanarayana, 2006), speech enhancement (B. Yegnanarayana and S. R. Mahadeva Prasanna and K. Sreenivasa Rao, 2002), emotion analysis (Gangamohan et al., 2013), etc. Hence, in this study, our focus is on the processing of high SNR regions of speech.

The remaining part of the paper is organized as follows: Section 2 describes the basis for the present study. Databases used and feature extraction procedures are described in Section 3. In Sections 4 and 5, description of the AANN models for capturing the excitation source and vocal tract system information are given. Emotion detection experiments and discussion on results are given in Section 6. Finally, Section 7 gives a summary and scope for further study.

2 Basis for the Present Study

Speech production characteristics are changed while producing emotional speech, and the changes are mostly in the excitation component. The changes are not sustainable for longer periods, and hence are not likely to be present throughout. This is due to an extra effort needed to produce the emotional speech. The primary effect is on the source of excitation due to pressure from the lungs and the vibration of the vocal folds. Moreover, the changes in production can be affected only in some selected voiced sounds. Hence, some neutral speech segments are also present in emotional speech. While most changes from perception point of view take place at the suprasegmental level (pitch, intensity and duration), it is less likely that significant changes take place at the segmental level (vocal tract resonances). Changes at the suprasegmental level are mostly learnt features (acquired over a period of time). It is difficult to find consistent suprasegmental features which can form a separate group for each emotion. In this study, changes in the subsegmental features are examined for discriminating neutral and emotional speech of a speaker (speaker-specific) using AANN models. The features are referred to as subsegmental features, since we consider only 1-5 ms around the glottal closure of the voiced excitation for deriving these features.

3 Emotion Speech Databases and Feature Extraction

Two types of databases (semi-natural and simulated) are used for discrimination of neutral and emotional speech.

3.1 Emotion Speech Databases

Semi-natural database was collected from 2 female and 5 male speakers of Telugu language. They are uttered in 4 emotions (angry, happy, neutral and sad), and it was named as IIIT-H Telugu emotion database (Gangamohan et al., 2013). Speakers were asked to script the text themselves by remembering past memories and situations which make them emotional. The lexical content is different for each speaker and for each emotion. The data was collected in 2 or 3 sessions for each speaker, and consists of around 200 utterances. The complete database was evaluated in perceptual listening tests for recognizability of emotions by 10 listeners. A total of 130 utterances were selected, in which anger, happy, neutral and sad are 35, 27, 34 and 34 utterances, respectively.

To test the effectiveness of language independent emotion detection, the Berlin emotion speech database (EMO-DB) (Burkhardt et al., 2005) is chosen. Ten professional native German actors (5 male and 5 female) were asked to speak 10 sentences (emotionally neutral sentences) in 7 different emotions, namely, anger, happy, neutral, sad, fear, disgust and boredom in one or more sessions. The total database was evaluated in a perception test by 20 listeners regarding the recognizability of emotions that had recognition rate better than 80% and naturalness better than 60% for analysis. A total of 535 good utterances were selected, in which anger, happy, neutral, sad, fear, disgust and boredom are 127, 71, 79, 62, 69, 46 and 81 utterances, respectively.

3.2 Feature Extraction

The features related to the excitation source and the vocal tract system components of speech signal are used in this study. The major source of excitation of the vocal tract system is due to vocal folds vibration at the glottis. The instant of significant excitation is due to sharp closure of the vocal folds, and it is almost like impulse. Hence, the high SNR of speech is present around Glottal Closure Instants (GCIs). By extracting the GCIs from the signal, it is possible to focus the analysis around the GCIs to further extract the information from both the excitation source and the vocal tract system. The features investigated for the detection of emotions are Linear Prediction (LP) residual for excitation source and weighted Linear Prediction Cepstral Coefficients (wLPCCs) for vocal tract system component extracted around the GCIs of speech signal. For this purpose, we use two signal processing methods, one is, a recently proposed method of Zero Frequency Filtering (ZFF) (Murty and Yegnanarayana, 2008) for extraction of GCIs, and another is LP analysis (Makhoul, 1975) for extraction of LP residual and wLPCCs.

3.2.1 Zero Frequency Filtering (ZFF) Method

The motivation behind this study was that, the effect of impulse-like excitation is reflected across all frequencies including zero frequency (0 Hz) of the speech signal. The method uses the zero frequency filtered signal obtained from the speech signal by filtering the signal through a cascade

of two 0 Hz resonators to get the epoch locations. The instants of negative-to-positive zero crossings (NPZCs) of the ZFF signal correspond to the instants of significant excitation, i.e., epochs or Glottal closure instants (GCIs) in voiced speech (Murty and Yegnanarayana, 2008). This method is also useful for detecting voiced and unvoiced regions. Because of significant contribution by the impulse-like excitation, ZFF signal energy is high in voiced regions (Dhananjaya and Yegnanarayana, 2010). In this paper, we considered only voiced regions for analysis.

3.2.2 Linear Prediction (LP) Analysis

The production characteristics of speech has the role of both excitation source and the vocal tract system. LP analysis with proper LP order gives the excitation source (LP residual) component and vocal tract system component through LPCs. In the LP residual, the region around the GCI within each pitch period is used for processing the high SNR regions of speech (Reddy et al., 2010). For deriving the LP residual and LPCs, a 10^{th} order LP analysis is used on the signal sampled at 8 kHz. Two pitch periods of signal are chosen for deriving the residual and a 4 ms segment (i.e, 32 samples) of the LP residual is chosen around each epoch to extract the information from the excitation source component. The vocal tract system characteristics around each GCI is represented by a 15 wLPCC vector derived from the 10 LPCs.

4 AANN Models for Capturing the Excitation Source Information

Autoassociative Neural Network (AANN) is a feedforward neural network model which performs identity mapping (Yegnanarayana and Kishore, 2002; Yegnanarayana, 1999). Once the AANN model is trained, it should be able to reproduce the input at the output with minimum error, if the input is from the same system. The AANN model consists of one input layer, one or more hidden layers and one output layer (Haykin, 1999). The units in the input and output layers are linear, whereas the units in the hidden layers are nonlinear. The AANN is expected to capture the information specific to the neutral speech present in the samples of LP residual. A five layer neural network architecture (Fig. 1) is considered for the study.

The structure of the network $33L\ 80N\ xN\ 80N\ 33L$, is chosen for extracting the neutral speech information using the 4 ms LP residual around each GCI. Here L refers to linear units, and N refers to nonlinear ($tanh()$) output function of units. Here x refers to the number of units in the compression layer, which is varied to study its effect on the model's ability to capture the neutral speech specific information. The sizes of input and the output layer are fixed by the number of residual samples (around each GCI) used to train and test the models. The hidden layers provide flexibility for mapping and compression.

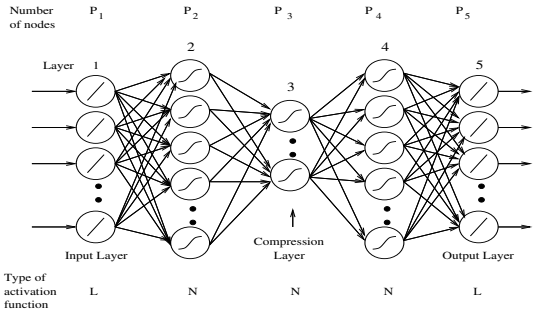

Figure 1: Structure of the AANN model

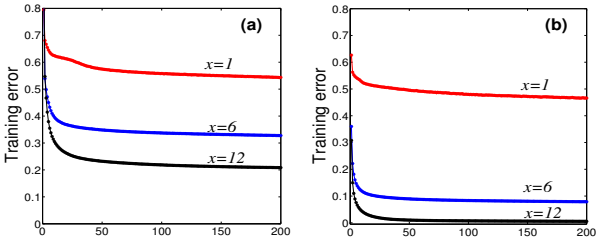

Figure 2: Training error as a function of iteration number, for (a) excitation source models and (b) vocal tract system models. Here x indicates number of nodes in the compression layer.

The network is trained for 200 iterations. The training error plots are shown in Fig. 2(a) for different values of the number of units (x) in the compression layer. From Fig. 2(a), it is observed that the error is decreasing with number of iterations, and hence the network is able to capture neutral speech information of a speaker in the residual. It can also be observed that the decrease in error is more as the number of units in the compression layers are increasing. Even if the error decreases, the generalizing ability may be poor beyond a certain limit on the number of units in the compression layer (Reddy et al., 2010).

5 AANN Models for Capturing the Vocal Tract System Information

A 5-layer AANN model with the structure $15L\ 40N\ xN\ 40N\ 15L$ is used for extracting the neutral speech specific information using 15 dimensional wLPCC vectors derived using LP analysis on two pitch period segment around each GCI. The model is expected to capture the distribution of the feature vectors of neutral speech of a speaker. The training error plots for a neutral speech of a speaker for x =1, 6 and 12 units in the compression layer are shown in Fig. 2(b). It is observed that the information in the distribution of the feature vectors is captured. The ability of the model to capture the neutral speech information can be determined through emotion detection experiments, as described in Sec. 6.

6 Emotion Detection Experiments

In order to know the capturing ability of AANN models for emotion detection, in the experiments we used all the speech samples from two types of databases described in Sec. 3.1. The speech samples are picked randomly while training and it is noted that the experiments are conducted in lexical independent way. For EMO-DB database, a universal background model (UBM) is built from 10 speakers (5 male and 5 female) using 15 s neutral speech data from each speaker. We have used all 10 speakers data for emotion detection experiments. Approximately 20 s of neutral speech data from a speaker is used to train over the UBM to build the speaker-specific neutral speech AANN models using 200 iterations. For testing, emotional speech utterance is presented to the neutral speech AANN model, and the mean squared error between the output and input, normalized with the magnitude of the input, is computed.

Fig. 3(a) shows the plots of the normalized errors obtained from the neutral speech AANN models using excitation source information (LP residual) of a speaker at each GCI. The solid ('—') line is the output from the model of the neutral speech of the same speaker. The emotional speech test utterance is fed to the neutral speech AANN models and the resulting error is shown by dotted ('$\cdots$') lines. The plots correspond to three different cases, i.e., for 1, 6, 12 units in the middle compression layer. It can be seen that the solid line (neutral speech) has the lowest normalized error values for most of the frames (from GCI index 1 to 150). As the number of units in the middle layer increases, the error for the neutral speech is decreasing and the error for emotional speech is increasing. Similar observations can be made from Fig. 3(b) for the error plots for an emotional test utterance tested against neutral speech models using vocal tract system information (wLPCCs).

Since the neutral speech AANN models are built, it is expected that the error range should show discrimination for neutral and emotional speech. It is observed that the network is giving lower error values when the test utterance is neutral and higher error values when the test utterance is emotional. Using a threshold on the averaged normalized error value (averaged over all the frames of an utterance), emotion detection studies are performed. The averaged normalized error is given by

$$\frac{1}{l}\sum_{i=1}^{l}\frac{\|y_i - z_i\|^2}{\|y_i\|^2}. \qquad (1)$$

where y_i is the input feature vector of the model, z_i is the output given by the model, and l is the number of frames of the test utterance.

The results of emotion detection (neutral vs emotion) using the excitation source information and the vocal tract system information for EMO-DB are shown in the Table 1. To test the effectiveness of language independent emotion detection, similar study is performed on IIIT-H Telugu emotion database, and the results are shown in Table 2. The accuracy for EMO-DB database using excitation source information is 91%, which is nearly 8% higher than that for the vocal tract system information. This is because, for the production of emotional speech, the primary effect is on the excitation source component such as pressure from the lungs and the changes in the vocal fold vibration. Similar observations can be made for the IIIT-H Telugu emotion database. For both the databases, it is indicative that the excitation source information carries more information than the vocal tract system information, and also the performance is consistent across the speakers. It is observed that, proposed excitation source and vocal tract system features with AANN models provides an improvement of approximately 10% and 3% over the recently proposed emotion detection method (Arias et al., 2014; Arias et al., 2013) which uses the energy and pitch contour modeling

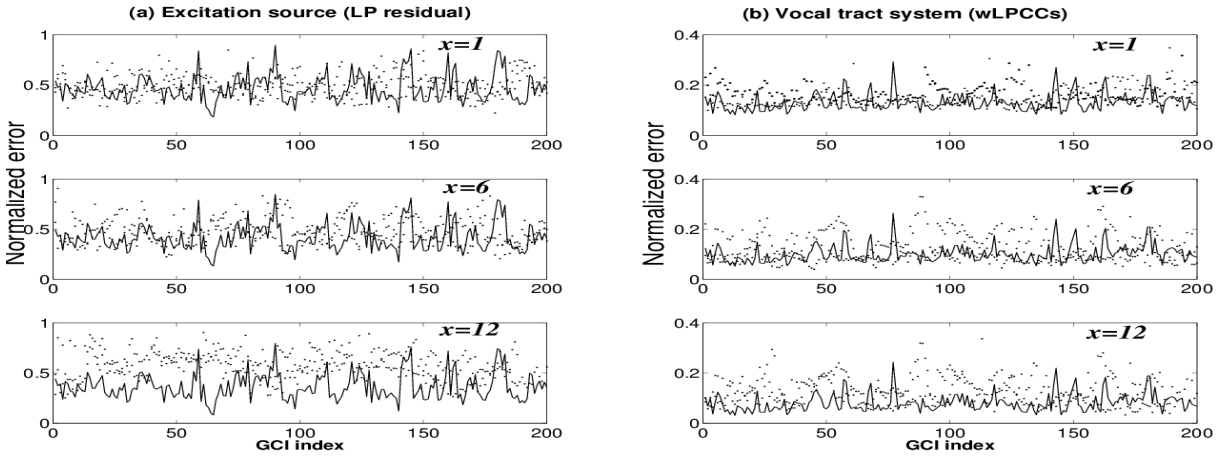

Figure 3: Normalized errors obtained from AANN models of various architectures, using (a) excitation source (b) vocal tract system information. In each plot, solid line ('—') and dotted line ('· · · ') correspond to neutral and emotion utterance normalized error curves, respectively.

with functional data analysis (accuracy is 80.4%) for EMO-DB database.

From Tables 1 and 2, it is observed that the higher activated emotion states (anger, happy, disgust and fear) are more discriminative compared to the lower activated emotion (sad and boredom) states. This is in conformity with the studies reported in (Jeon et al., 2011; Lee et al., 2011; Shami and Verhelst, 2007), where generally the acoustic features effectively discriminate between high activated emotions and low activated emotions. The accuracies for lower activation states is low, because some of the speech segments in these emotion signals are closer to neutral speech. Thus, AANN models are able to discriminate higher activated and lower activated states using neutral speech reference model. This study can be extended by training models with the lower activation states and testing with the higher states and vice versa. It is also noted that (from Fig. 3), all the frames of an utterance are not important in taking the decision, i.e., emotion information may not be uniformly distributed in time and hence, automatic usage of the high confidence frames may improve the accuracy. Also, the proposed emotion detection system can be evaluated using other spectral parameters such as variants of LPCs, MFCCs, MSFs (Schuller et al., 2013; Schuller et al., 2011; Schuller et al., 2009; Zeng et al., 2009; Ayadi et al., 2011; Eyben et al., 2012) etc. The advantages of the present study are, it is independent of lexical content used, language and channel.

Table 1: *Results for neutral vs emotion detection for Berlin EMO-DB database (in percentage).*

	Excitation Source	**Vocal Tract System**
Neutral vs Anger	100	100
Neutral vs Happy	100	100
Neutral vs Sad	73.50	62.93
Neutral vs Disgust	100	81.24
Neutral vs Fear	100	91.32
Neutral vs Boredom	72.72	64.02
Neutral vs Emotion	**91.03**	**83.25**

Table 2: *Results for neutral vs emotion detection for IIIT-H Telugu emotion database (in percentage).*

	Excitation Source	**Vocal Tract System**
Neutral vs Anger	100	92.86
Neutral vs Happy	100	96.43
Neutral vs Sad	82	75.4
Neutral vs Emotion	**94**	**88.23**

7 Summary

In this paper, we have demonstrated the significance of pitch synchronous analysis of speech data using AANN models for discriminating neutral and emotional speech. We have shown that the excitation source information of neutral speech is captured using 4 ms LP residual around the GCI, and the vocal tract system information of neutral speech is captured using 15 dimensional wLPCC vectors derived from 10 LPCs around each GCI. Emotion detection (neutral vs emotion) experiments were conducted using two databases one is IIIT-H Telugu emotion database, and the other is a well known emotion speech database EMO-DB. The results show that excitation source component

carries more information than that of the vocal tract system. Further, it can be extended for the discrimination among the emotions, such as discrimination of anger and happy by training anger models and testing with happy emotion utterances, and vice versa. Also, it is important to develop speaker-specific models by determining suitable AANN models for individual speakers and emotions. It is also necessary to explore methods to combine the evidence from excitation source and vocal tract system for emotion detection.

References

J.P. Arias, C. Busso, and N.B. Yoma. 2013. Energy and F0 contour modeling with functional data analysis for emotional speech detection. In *INTERSPEECH*, pages 2871–2875, August.

Juan Pablo Arias, Carlos Busso, and Nestor Becerra Yoma. 2014. Shape-based modeling of the fundamental frequency contour for emotion detection in speech. *Computer Speech and Language*, 28(1):278 – 294.

Paeschke Astrid and W F Sendlmeier. 2010. Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements. In *SpeechEmotion*, pages 75–80.

Theologos Athanaselis, Stelios Bakamidis, Ioannis Dologlou, Roddy Cowie, Ellen Douglas-Cowie, and Cate Cox. 2005. Asr for emotional speech: Clarifying the issues and enhancing performance. *Neural Networks*, 18(4):437–444.

Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572 – 587.

B. Yegnanarayana and S. R. Mahadeva Prasanna and K. Sreenivasa Rao. 2002. Speech enhancement using excitation source information. In *ICASSP*, volume 1, pages 541–544.

A. Bajpai and B. Yegnanarayana. 2008. Combining evidence from subsegmental and segmental features for audio clip classification. In *TENCON IEEE Conference*, pages 1–5, Nov.

Murtaza Bulut and Shrikanth Narayanan. 2008. On the robustness of overall f0-only modifications to the perception of emotions in speech. *J. Acoust. Soc. Am.*, 123(6):4547–4558, June.

Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. 2005. A database of german emotional speech. In *INTERSPEECH*, pages 1517–1520.

C. Busso, S. Lee, and S.S. Narayanan. 2007. Using neutral speech models for emotional speech analysis. In *INTERSPEECH*, pages 2225–2228, August.

Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. 2009. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech & Language Processing*, 17(4):582–596.

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

N. Dhananjaya and B. Yegnanarayana. 2010. Voiced/nonvoiced detection based on robustness of voiced epochs. *IEEE Signal Processing Letters*, 17(3):273–276, March.

Florian Eyben, Anton Batliner, and Björn Schuller. 2012. Towards a standard set of acoustic features for the processing of emotion in speech. *Proceedings of Meetings on Acoustics*, 16.

P Gangamohan, Sudarsana Reddy Kadiri, and B. Yegnanarayana. 2013. Analysis of emotional speech at subsegmental level. In *INTERSPEECH*, pages 1916–1920, August.

Ali Hassan and Robert I. Damper. 2012. Classification of emotional speech using 3dec hierarchical classifier. *Speech Communication*, 54(7):903–916.

Simon Haykin. 1999. *Neural networks: A Comprehensive Foundation*. Prentice-Hall International, New Jersey, USA.

Je Hun Jeon, Rui Xia, and Yang Liu. 2011. Sentence level emotion recognition based on decisions from subsentence segments. In *ICASSP*, pages 4940–4943.

Chul Min Lee and Shrikanth S. Narayanan. 2005. Toward detecting emotions in spoken dialogs. *IEEE Trans. Audio, Speech, and Language Processing*, 13(2):293–303.

Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. 2011. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9-10):1162–1171.

Iker Luengo, Eva Navas, and Inmaculada Hernáez. 2010. Feature analysis and evaluation for automatic emotion identification in speech. *IEEE Transactions on Multimedia*, 12(6):490–501.

J. Makhoul. 1975. Linear prediction: A tutorial review. *Proc. IEEE*, 63:561–580, Apr.

Leena Mary and B. Yegnanarayana. 2008. Extraction and representation of prosodic features for language and speaker recognition. *Speech Communication*, 50(10):782–796.

Marc Mehu and Klaus R. Scherer. 2012. A psycho-ethological approach to social signal processing. *Cognitive Processing*, 13(2).

Donn Morrison, Ruili Wang, and Liyanage C. De Silva. 2007. Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, 49(2):98–112.

Iain R. Murray and John L. Arnott. 1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *J. Acoust. Soc. Am.*, 93(2):1097–1108.

K. Sri Rama Murty and B. Yegnanarayana. 2006. Combining evidence from residual phase and MFCC features for speaker recognition. *IEEE Signal Process. Letters*, 13(1):52–55, Jan.

K. Sri Rama Murty and B. Yegnanarayana. 2008. Epoch extraction from speech signals. *IEEE Trans. Audio, Speech, and Language Processing*, 16(8):1602–1613, Nov.

Sri Harish Reddy, Kishore Prahallad, Suryakanth V. Gangashetty, and B. Yegnanarayana. 2010. Significance of pitch synchronous analysis for speaker recognition using aann models. In *INTERSPEECH*, pages 669–672.

Klaus R. Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2):227–256.

Björn Schuller, Stefan Steidl, and Anton Batliner. 2009. The interspeech 2009 emotion challenge. In *INTERSPEECH*, pages 312–315.

Björn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wöllmer, André Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll. 2010. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Tran. Affective Computing*, 1(2):119–131.

Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9-10):1062–1087.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A. Müller, and Shrikanth Narayanan. 2013. Paralinguistics in speech and language - state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39.

A. Shahina and B. Yegnanarayana. 2007. Mapping speech spectra from throat microphone to close-speaking microphone: A neural network approach. *EURASIP J. Adv. Sig. Proc.*, 2007.

Mohammad Shami and Werner Verhelst. 2007. An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech Communication*, 49(3):201–212.

Carl E. Williams and Kenneth N. Stevens. 1972. Emotions and speech: Some acoustical correlates. *J. Acoust. Soc. Am.*, 52(4B):1238–1250.

B. Yegnanarayana and S. P. Kishore. 2002. AANN - an alternative to GMM for pattern recognition. *Neural Networks*, 15:459–469, Apr.

B. Yegnanarayana, K. Sharat Reddy, and Kishore Prahallad. 2001. Source and system features for speaker recognition using AANN models. In *Proc. Int. Conf. Acoustics Speech and Signal Processing*, pages 491–494, Salt Lake City, Utah, USA, May.

B. Yegnanarayana. 1999. *Artificial Neural Networks*. Prentice-Hall of India, New Delhi.

Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(1):39–58.