TOWARDS USING PROSODY IN SPEECH RECOGNITION/UNDERSTANDING SYSTEMS: DIFFERENCES BETWEEN READ AND SPONTANEOUS SPEECH

Kim E.A. Silverman, Eleonora Blaauw[1], Judith Spitz, John F Pitrelli

Speech Technology Group
Artificial Intelligence Laboratory
NYNEX Science and Technology
White Plains, NY, 10604, U.S.A.

1. Ph.D. student in Institute for Language and Speech, University of Utrecht

1. ABSTRACT

A persistent problem for keyword-driven speech recognition systems is that users often embed the to-be-recognized words or phrases in longer utterances. The recognizer needs to locate the relevant sections of the speech signal and ignore extraneous words. Prosody might provide an extra source of information to help locate target words embedded in other speech. In this paper we examine some prosodic characteristics of 160 such utterances and compare matched read and spontaneous versions. Half of the utterances are from a corpus of spontaneous answers to requests for the name of a city, recorded from calls to Directory Assistance Operators. The other half are the same word strings read by volunteers attempting to model the real dialogue. Results show a consistent pattern across both sets of data: embedded city names almost always bear nuclear pitch accents and are in their own intonational phrases. However the distributions of tonal make-up of these prosodic features differ markedly in read versus spontaneous speech, implying that if algorithms that exploit these prosodic regularities are trained on read speech, then the probabilities are likely to be incorrect models of real user speech.

2. INTRODUCTION

This work addresses two related questions. One is whether spontaneous goal-directed utterances collected from real users in a particular application domain exhibit reliable prosodic patterns that could be exploited by recognition algorithms. We focus on to-be-recognized words that are spoken within longer utterances, in order to investigate whether these embedded words have particular prosodic characteristics that could help a recognizer to locate them. One of the original motivations for this study was our observation from informal listening to our corpus that such embedded words bear nuclear pitch accents. If this is a consistent pattern, it would mean that they are (1) louder, longer and more clearly articulated than they would be without nuclear accents, and (2) they would bear characteristic fundamental frequency movements.

Corpora of spontaneous goal-directed speech from real users are not readily obtainable, and so it is common practice to record speech read out by volunteers in order to develop, train and test recognition algorithms. To the extent that the prosody of read speech differs from that of spontaneous goal-directed speech, such "laboratory" corpora may obscure or misrepresent any reliable prosodic properties found in spontaneous "real user" speech. Consequently the second question investigated in this work is whether such patterns can also be found in recordings of speech read out by volunteers. We are interested in prosodic differences between read and spontaneous speech because of their relevance to speech recognition, and for increasing the naturalness of synthetic speech.

It is worth pointing out a methodological issue at this stage: the prosody used when people are reading can of course differ dramatically from that used in spontaneous communication. Speech databases that we know of vary widely in how much effort was taken to ensure that the prosody of the speech realistically reflects the speech that recognizers have to deal with in real-world applications. In this experiment we chose to do everything we could to encourage our volunteers to use realistic spontaneous-sounding prosody. Most existing speech corpora used in the recognition field have been collected with less emphasis on realistic prosody. We therefore believe that the read speech in this experiment is as similar as possible to spontaneously produced utterances. The degree of prosodic similarity that we report between read and spontaneous speech represent a "best case".

3. DISCOURSE DOMAIN

Our particular application is automatic recognition of the name of a city in telephone calls to a Directory Assistance Operator. A corpus of 26,946 recordings of real users has been collected and was reported on in detail in a previous DARPA meeting[1]. In each case a caller was played an automated request for the name of a city. In 37% of those utterances that do contain a city name, that city name is embedded in a longer string of connected speech. In some cases there is relatively little extraneous material (*"In Boston, please"*), but often (55% in this corpus) there is considerably more (*"Yes Ma'm Central Auto Service in*

Stoughton, please."). We refer to these as "complex embedded" utterances. In all of these embedded utterances speakers have considerably more options concerning how they say the speech than when they say only an isolated city name, and so we would expect prosody to contribute significantly to the variability in the signal. The current study deals with the complex embedded utterances, because (i) these represent the more serious challenge for speech recognition, and (ii) these contain richer prosodic variation that is more representative of spontaneous speech in other discourse domains.

Although the current study focuses on a telephone network application, we believe that the results have general applicability to the behavior of users of spoken language systems (SLS). Often users will answer a request for a single item of information with a sentence containing not only the requested item, but also (1) extraneous material, and (2) answers to anticipated subsequent requests in the discourse. For example, in an ATIS-like domain, a user may answer a request for a destination with "I'd like to arrive in Boston on Tuesday morning before 9:30 am", bundling the arrival day and arrival time into the same utterance. Prosody can mark all of the discourse-relevant information-bearing words in such an answer, and so could help a SLS to avoid reprompting for material that has already been said.

For this investigation, we selected 80 of the spontaneous complex embedded utterances that reflected the variation in length and structure of the larger set. Each utterance was spoken by a different speaker, half of the speakers were male and half were female. The shortest utterance was two words (*"Arlington, McCarthy"*) and the longest was twenty words (*"Have you a listing in Jamaica Plain for Robert Scheinkopf - S - C - H - E - I - N - K - O - P - F"*). Half of the 80 utterances were in "telegram" style - bearing few or no function words (*"Boston Woolworth's on Washington Street"*), and the set was chosen to reflect variation in whether the target city name was a first, medial, or last content word in the utterance.

We then collected a matched corpus in which volunteers called an automated recording facility in our laboratory and read out orthographic transcriptions of the same utterances. Participants knew that these texts were originally spoken by people calling Directory Assistance asking for information about a telephone number, and were encouraged to rehearse the items several times before calling with this in mind. Participants confirmed during subsequent debriefing that they had tried to make their utterances sound realistically natural, acting out the situation of making a telephone call to get information. Each participant read a list of 25 sentences: the first and last two were fillers, and one of the middle sentences was the utterance relevant to the current study. This "read speech" corpus consists of 2000 utterances altogether, collected from 80 volunteers, of which one utterance per reader is used in this investigation. The reader of each utterance was of the same sex as the speaker of the spontaneous version.

4. PROSODIC ANALYSIS METHODS

We use "prosody" to refer to the acoustic/phonetic bracketing structure, locations of boundaries, and the choice and distribution of tonal features. This suprasegmental organization has been shown to affect not only duration and fundamental frequency but also such phenomena as coarticulation, devoicing, laryngealization, and allophonic variation. Thus it is a potential information source for factoring out variability in acoustic-phonetic models, locating word boundaries, disambiguating alternative parses or interpretations, and locating embedded keywords. In this study we focus primarily on the last of these.

The prosody in the read and spontaneous versions of the utterances was manually transcribed by two people via interactive listening and graphical displays of the speech waveform, total signal energy, and extracted fundamental frequency contour. This signal-processing and display was performed with version 2.0 of the WAVES+ software package[4]. Each transcriber labelled an overlapping subset of the utterances, enabling us to compare their transcriptions for almost half of the corpus. In addition, a number of direct measurements were taken from the acoustic signals.

4.1. Prosodic Transcription Scheme

The utterances were transcribed using the draft prosodic transcription conventions developed at the First Prosodic Transcription Workshop hosted by the Spoken Language Systems Group in MIT in August 1991. Briefly, this is a set of labels for the tonal (pitch) structure and boundary structure of spoken American English. The tonal labels are a subset of Pierrehumbert's model[2]: this approach views pitch contours as composed of a sparse linear sequence of accents and tones selected from a relatively small inventory. In the draft scheme used below, the inventory of pitch accents is reduced to **H***, **L***, **L+H***, **L*+H**, and the downstep feature is marked explicitly (e.g. **H*** versus **!H***). Lack of confidence is marked by affixing a "?" after the symbol. Boundaries are a subset of the break indices used by Price et al [3]. The labelling process consisted of locating and identifying the pitch accents, phrase accents, and boundary tones, and assigning a strength to each inter-word boundary (from **0** => cliticized word; to **4** => full intonational phrase boundary). A transcriber can affix a + or - to indicate uncertainty about the strength of a boundary.

One of the transcribers received one day's training beforehand, and supplemented that by reading portions of Pierrehumbert [2] and Silverman [5]. The other transcriber received about a half day's training. Both transcribers would occasionally consult with the first author concerning particularly unclear phenomena.

5. RESULTS

5.1. Reliability Across Transcribers

One of the stated aims of the First Prosodic Transcription Workshop was that the transcription conventions should be easily taught, and that different transcribers should agree at least 80% of the time.

We could test this in this experiment, because 36 of the spontaneous and read versions (i.e. 72 utterances in all) were labelled by both transcribers. Because transcriptions are linear strings of symbols, one way to calculate agreement between 2 transcribers is:

$$Agreement = 100\left(\frac{Matches}{Matches + Insertions + Substitutions}\right)$$

where:

Matches = number of symbols in the string where the transcribers agree concerning location and the symbol itself[1], *Insertions* = number of symbols marked by one transcriber only (an omission by either transcriber is equivalent to an insertion by the other), and *Substitutions* = number of locations where each transcriber used a different symbol.

Table 1 shows the agreement separately calculated for the tonal and boundary transcriptions under two criteria. Overall the agreement is quite satisfactory. *Exact match* means both transcribers had to use exactly the same symbols in the same locations to score a match. *Near match* slightly relaxes the criteria for matching in the following ways:

Near tonal match: (1) phrase-initial **H*** matches phrase-initial **L+H***; (2) a **H*** or **L+H*** match the corresponding downstepped variants of themselves (**!H*** and **!L+H***, respectively), (3) an accent matches its uncertain variant (e.g. **H*** matches **H*?**)

Near boundary match: (1) a **0** (= cliticized word) matches a **1** (= normal phrase-medial interword boundary), (2) a **1** matches a **2** (= separation between words, but with no tonal correlates of a boundary)

If agreement includes near-matches, then we have clearly met the reliability criteria. If not, then we still have met it in the tonal transcriptions, but not in the boundary transcriptions. Most of the disagreements concerned whether some word boundaries were cliticized (e.g. between the first and second words in "*Could I have the...*" versus "*C'd I have the...*"). In the subsequent preliminary analyses of the tonal transcriptions, we used the more experienced transcriber's decisions in cases where there is disagreement.

1. Final boundaries at the right-hand edge of utterances are excluded from this analysis because they would artificially inflate the agreement scores: both transcribers agreed 100% that all utterances ended with a **4** boundary.

	Exact Match	Near Match
Tonal Structure	81%	92%
Boundary Structure	68%	94%

Table 1: Percent agreement between transcribers on tonal and boundary structure.

5.2. Comparison of Read and Spontaneous Versions: Intonation

Our initial informal impression which motivated this study was borne out by the transcriptions: in both corpora the embedded city names usually bear a nuclear accent (94% of spontaneous, 97% of read utterances, no significant differences), and are set off in an intonational phrase of their own. Moreover the tonal combinations carried by these city names represent only a relatively small subset of the possible combinations that can occur in spoken English. Within the transcription framework used in this study, there are 16 different possible combinations of pitch accent, phrase accent and boundary tone (the three tonal elements of a city name in most of our corpus). However, the only five that actually occurred on the city names were:

Pitch accent	Phrase accent	Final boundary tone[a]
H*	**L**	**L]**
L+H*	**L**	**L]**
L*	**H**	**H]**
H*	**H**	**H]**
L+H*	**H**	**H]**

a. To avoid confusion with (i) the results expressed below as percentages, and (ii) initial boundary tones in Pierrehumbert [2], we follow the convention in Silverman [5] of using "]" instead of "%" for final boundary tones.

The first two of these tunes are falls, the last three are types of rises. These same five tunes occurred in both the read and the spontaneous corpora. We interpret this as another similarity between the read and spontaneous corpora: the readers not only succeeded in putting nuclear accents on the city names, but also chose from the same inventory that is used in spontaneous interactions in this domain. However

although the embedded city names were almost all nuclear in the read and spontaneous utterances, the distributions of the five tunes across the corpora were not at all the same. It is commonplace in the literature to categorize nuclear pitch very grossly movements into rising (or high level) versus falling. This corresponds in our case largely to the phrase accent being **H** or **L** (this would not be the case if there had been any **L** phrase accents followed by **H]** boundary tones). In Table 2 we compare the read and spontaneous versions of 79[1] of the pairs of city names according to this gross division. For the few city names that bore prenuclear **H*** accents, we categorized them as falling if the next accent was either a **L*** or downstepped, else as rising.

		Read version		
		Rising	Falling	
Spontaneous version	Rising	27%	47%	74%
	Falling	8%	19%	27%
		34%	66%	

Table 2: Agreement between spontaneous and read versions of each city name. All percentages are out of the 79 pairs included in this analysis

One common view of prosody is that it is determined by syntax, that there is a default prosody for any given sentence which is derivable from the word string itself. If this is true, or if the way city names are read out resembles how they are spoken spontaneously, then the data should be concentrated in the upper left and lower right cells of Table 2. In fact, less than half of the data (46%) lies in these two cells. The main reason is that 47% of the city names were spoken with a rising intonation in spontaneous versions, but with falling intonation in the read versions.

This shift from rising to falling intonation is also reflected in the marginal totals: 73% of the spontaneous city names had rises, but only 34% of their read counterparts did. The data argue that prosody is not directly derivable from the word string itself. Two possible reasons for this difference are:

- A rising intonation is a marker of politeness in this particular dialogue context. When volunteers participate in a recording session, even when they attempt to act out the real dialogue, they do not feel compelled to be polite to the recording equipment.
- In the real interaction with a telephone operator the speaker uses rising intonation to seek confirmation that the operator has indeed understood the city name. Speakers may not be consciously aware that they do this, and so fail to replicate it when attempting to emulate the interaction

The preponderance of falls in read speech, when compared to the preponderance of rises in spontaneous speech, has a number of implications for speech technology. One of these concerns the acoustic models in a recognizer: low final boundary tones (which are located at the right hand edge of most of the falling nuclear accents, and therefore are common in read speech) tend to be associated with laryngealization and devoicing of the segmental material. Consequently these spectral effects will be built into acoustic models that are trained on read speech, but will be the exception rather than the rule in the spontaneous speech that a recognizer will ultimately have to process.

These rising-versus-falling differences also bear on a potential use of prosody for speech understanding systems: in a system where the user can both ask questions and also deliver answers to questions asked by the system itself, the natural language processing part of the discourse manager could be helped if it could use prosody to distinguish between these two different speech acts. Suggestions have been made that questions usually have high or rising intonation, whereas information-delivering statements usually have falling intonation. The current results indicate that at least in the application domain used in this experiment, this distinction is more complicated. Users are delivering information in response to a question from the system, rather than asking the system a direct question themselves, and yet they use rising tunes more often than falling.

The tonal differences between the corpora were not restricted to the phrase accents alone. In read speech, 81% of the city names carried a **H***, whereas in spontaneous speech this was only 52%, with the remaining cases bearing either a **L+H*** (35%) or a **L*** (8%). The majority of the city names were final in their intonational phrase, and therefore contained an additional boundary tone on the right. In the spontaneous corpus, 76% of these were **H]** and 12% were **L]**. Once again, this order was reversed in the read corpus (28% were **H]** and 72% **L]**).

5.3. Comparison of Read and Spontaneous Versions: Pauses

The characteristics and distribution of pauses in these utterances also showed reliable patterns and important differences between read and spontaneous speech. The following summary is based on all pauses in the utterances, not just those around city names. Some pauses occurred at "grammatical" positions, as in:

"In Boston <...> may I have the number of..."

"...the number John Smith <...> in Boston",

others at "ungrammatical" positions:

1. One of the read versions was truncated in such a way that it had to be left out of this analysis

"yes the number of <...> John Smith in <...> Boston".

This classification of pause types is common in the literature. While it seems to have intuitive appeal, we believe that it may be more of a continuum than a clear category distinction. Ungrammatical pauses may be reinterpreted as merely being located at more embedded levels of bracketing in a syntactic structure than grammatical pauses. At least in some cases the labelling of a pause as grammatical or ungrammatical may be a consequence of the researcher's preferred syntactic theory. In the current study, 91% of the ungrammatical pauses were located after the preposition within a prepositional phrase.

Like O'Shaughnessy [6], we found that while some pauses are located at grammatical boundaries, others are not. But the ratio distinguished between the two speech modes: 45% of all pauses were "ungrammatical" in the spontaneous speech, but only 11% in the read speech. Unlike O'Shaughnessy, we found that in both corpora ungrammatical pauses were longer than grammatical ones. Silent pauses in grammatical locations were twice as long in spontaneous speech (mean 0.45 seconds, standard deviation 0.29) as in read speech (mean 0.21 seconds, standard deviation 0.15). In the read corpus there was less variability in pause duration (mean 0.23 seconds, standard deviation 0.17) than in the spontaneous speech (mean 0.45 seconds, standard deviation 0.29). 85% of the filled pauses were located at ungrammatical positions One striking difference between the corpora was that in the read versions there were no filled pauses at all. Moreover, in only 18% of the read utterances did the readers place pauses in the same places as they occurred in the spontaneous versions. All of these were grammatical boundaries (*"Cambridge <silence> I'm looking for Pizza Ring"*) which also carried full intonational phrase boundaries. All other differences consisted of either omitting ungrammatical pauses or inserting grammatical ones where the original speakers did not.

We believe that in the spontaneous speech the ungrammatical pauses, and perhaps also some of the grammatical ones, reflect the speakers' lexical access delay and mark for the listeners that the post-pausal words are not easily predictable (i.e. information-rich) and therefore "worth waiting for". In read speech there is no comparable lexical access because all the words are already laid out on the printed orthography, and consequently this component of the information structure is not marked in the readers' utterances.

5.4. Prosodic Characteristics of Other Words

Although we do not yet have quantitative analysis specific to non-target speech, we do notice two consistent prosodic patterns in the remaining parts of the utterances outside of the city names.

The first pattern is that content words that are not directly conveying discourse-relevant information either bear no accent at all, or at their most salient bear only pre-nuclear accents and are not set off in phrases by themselves. Examples include the parenthesized words in:

"Could I (please) have the (number) for Watertown Police"

"Cambridge I'm (looking) for Pizza Ring"

"(I'm) (trying) to (find) the (exchange) for Cape Cod".

The second pattern returns us to the issue raised earlier in this paper that users often anticipate what questions will be asked subsequently in the dialogue. In the Directory Assistance domain subsequent questions will be for the name, and if that is likely to be ambiguous then there will be a request for further disambiguating information. The consistent behavior of users in our corpus is to mark this information in a similar way to how they mark the city name. Examples include:

"Quincy the Imperial Gardens on Sea Street"

"Yes I'd like the number of the Langley Deli in Newton please."

"Uh this is Quincy I'd like the number of the Quincy Police, not the emergency number of course."

One similarity is that these items tend to bear nuclear accents. But they differ in that these accents are often nuclear in an intermediate phrase, rather than a full intonational phrase. Thus they do not have the extra boundary tone, they exhibit less phrase-final lengthening, and are less likely to be followed by pauses.

Another common prosodic pattern arises when these extra compound nouns consist of more than one word, as illustrated in the above examples. Typically each word will bear a pitch accent, but they will not all be of the same type. The first accent is usually a L+H*, whereas subsequent ones are simple H*. That causes fundamental frequency to start low and rise to the first noun, and then stay high until the last one. Thereafter it moves into the phrase accent and is accompanied by lengthening of the material. In those cases where that phrase accent is L, then this contour appears to be an instance from American English of the "hat pattern" that has been described for British English and for Dutch [7]. Often in our spoken corpus the phrase accent is H. But in both cases, these patterns combine to somewhat set off the whole compound as a separate unit, in a way that could be exploited by a recognizer.

6. CONCLUSION

Read speech differs from spontaneous speech in some important ways: (i) although the tunes on focussed words are selected from the same inventory in both read and spontaneous speech, the prior probabilities of the tunes differ greatly -- spontaneous speech predominantly contains rises, read speech predominantly contains falls, (ii) pauses in read speech are shorter than in spontaneous speech, and they pre-

dominantly are located at structurally predictable positions (grammatical boundaries), whereas in spontaneous speech this generalization hardly holds true at all, (iii) read speech tends to not contain filled pauses. These differences argue that algorithms which are developed to exploit this information will need to be developed and trained on the basis of spontaneous speech from real users, rather than just from read speech.

These results are encouraging for locating embedded targets in speech recognition tasks: they show that when users respond to a query from an automated system, they mark the embedded information-bearing words with an acoustically-salient nuclear pitch accent and often precede and/or follow them by a pause.

For speech synthesis in the context of spoken language systems, these results suggest that listeners will better be able to understand and interpret synthesized utterances if the focussed information that they contain is (i) bears a nuclear tune, and (ii) is preceded by some lengthening of the immediately-preceding material and perhaps even the insertion of a short pause. Further investigations will address prediction of the tonal makeup of these patterns.

Acknowledgments

Sheri Walzman learned prosodic transcription and labored long doing careful labelling. Lisa Russell developed the automated recording facility, helped find suitable volunteers, and imposed organization and order on the data collection effort. Without the help of these two people this work would never have seen the light of day. Any abuses of their work nevertheless remain our own responsibility.

REFERENCES

1. Spitz, J. and the Artificial Intelligence Speech Technology Group. Collection and Analysis of Data from Real Users: Implications for Speech Recognition/Understanding Systems. *Proceedings of the Fourth DARPA Speech and Natural Language Workshop,* 1991.

2. Pierrehumbert, J. B., *The Phonology and Phonetics of English Intonation.* Ph.D. Dissertation, MIT 1980 (Distributed by Indiana University Linguistics Club, 1987).

3. Price, P., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, C. The Use of Prosody in Syntactic Disambiguation. *Journal of the Acoustical Society of America,* **90**, 6, pp 2956-2970, 1991.

4. Talkin. D. Looking at Speech. *Speech Technology,* **4**, 4, 1989.

5. Silverman, K. E. A., *The Structure and Processing of Fundamental Frequency Contours.* Ph.D. Dissertation, Cambridge University, 1987

6. O'Shaughnessy, D. Labelling Hesitation Phenomena in Spontaneous Speech. *Proceedings of the 1991 IEEE Workshop on Automatic Speech Recognition,* Arden House, 1991.

7. de Pijper, J. R., *Modelling British Intonation.* Foris: Dordrecht, 1983.