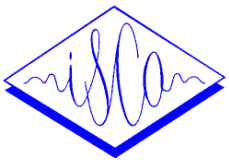

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

STATISTICAL AND LINGUISTIC ANALYSES OF F_0 IN READ AND SPONTANEOUS SPEECH[1]

Nancy A. Daly and Victor W. Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 U.S.A.

10.21437/ICSLP.1992-233

ABSTRACT

This paper describes our study of prosodic differences between read and spontaneous speech taken from human/machine problem-solving dialogues, as captured by F_0. Our research had two goals: first, to show that significant intonational differences exist between speech from these two styles and that these differences can be expressed quantitatively, and second, to demonstrate that the encoding of prosodic information is more salient in spontaneous speech than in read speech. Our analysis of over 4000 read/spontaneous utterance pairs from many speakers indicates that the mean of F_0 is statistically significantly higher for spontaneous speech than for read, but that F_0 is about equally variable in the two styles. In addition, we found that the encoding of final boundary tone information is more easily obtained from spontaneous speech than from read speech. Over 80% of final boundary tones from spontaneous speech were correctly classified, as opposed to less than 70% of those from read speech.

INTRODUCTION

Prosody, the stress, rhythm, and intonation of speech, not only contributes greatly to the perceived naturalness of speech, but also conveys a great deal of lingustic information. Consequently, a considerable amount of research has been devoted to the understanding of prosodic encoding in the speech signal. However, most explicit use of prosody in speech technology, particularly intonation information, has been in the area of speech synthesis, where variability is not an issue.

Prosody can aid the process of speech recognition in many ways, spanning across different linguistic levels, including phonological, syntactic, and semantic ones. For example, lexical stress information can aid speech recognition by introducing stressed and unstressed allophones, and information about word stress patterns can aid lexical access [1, 2]. While such information is potentially of great use, it has seldom been explicitly incorporated into various systems due to a lack of understanding of the variability of its encoding across speakers, sentence types and tasks. Therefore, to properly model this variability, we must study a large corpus of speech taken from many speakers.

Prosodic information is helpful, and often essential, for speech understanding tasks in which the system must understand the *meaning* of the input utterance. For example, "I am *flying* to Chicago" and "I am flying to *Chicago*" convey entirely different emphases that can only be resolved by considering prosodic information. Human/machine dialogues involving interactive problem solving provide many examples of this type of prosodic encoding.

In these dialogues, speech is typically produced extemporaneously, one sentence at a time, in a goal-directed interactive fashion. To date, virtually all studies of prosodic features of natural speech have been performed on read speech. Read speech is commonly collected for speech analysis as it allows experimenters more control over data characteristics. However, it may not be an appropriate speech style to study for improved understanding of prosodic encoding in human/machine dialogues.

One example of some of the prosodic differences between spontaneous and read speech is seen in Figure 1. The figure shows F_0 contours for spontaneous and read versions of the sentence *Is Toscanini's better than Steve's* said by the same speaker. The first utterance was said spontaneously by the subject. About thirty minutes later, the subject was shown the text and asked to read it.

The F_0 contours of the two utterances differ noticeably from each other. F_0 is clearly higher on average in the spontaneous utterance than in the read utterance. The F_0 contours have different shapes as well: in the last voiced region of the sentence, spanning the vowel /iʸ/, F_0 rises in the spontaneous utterance, but falls in the read. The spontaneous version has a typical "query-rise" type intonation, while the read version has a "final-fall" type intonation, although both are queries. According to the intonation transcription convention proposed by Pierrehumbert [3], the former utterance would be marked as ending with a high boundary tone, the latter with a low boundary tone.

A casual inspection of this read/spontaneous utterance pair indicates that the prosody of each is greatly affected by speaking style. In their studies of larger bodies of read and spontaneous speech, several researchers have noted acoustic and linguistic differences between the two styles. In her work, Koopmans-van Beinum found several differences, especially in F_0, between read and spontaneous speech from a radio announcer's monologues [4, 5]. In addition, Howell and Kadi-Hanifi noted differences in pause incidence in read and spontaneous speech also taken from monologues [6]. These results lend credibility to the hypothesis about prosodic differences between speaking styles. Nevertheless, these results may not concur with those obtained with speech from human/machine dialogues, since the speech was taken from monologues.

[1]This research was supported by DARPA under Contract N00014-89-J-1332, monitored through the Office of Naval Research.

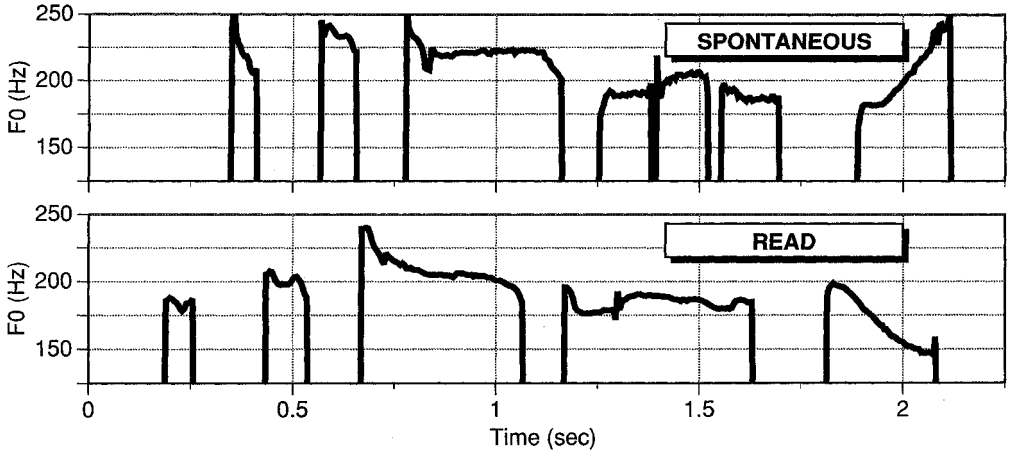

Figure 1: F_0 contours for read and spontaneous versions of the utterance *Is Toscanini's better than Steve's* said by the same speaker.

The goal of this paper is to understand and quantify intonation differences between read and spontaneous speech from human/machine dialogues. To reduce the scope of our study, we have focused on the encoding of intonation information in F_0 contours. We wish to use F_0 information to show general quantitative prosodic differences between speaking styles. In addition, as mentioned in our previous work [7], we believe that prosodic encoding is more salient in spontaneous communicative speech than in read. To demonstrate this, we will explore the differences in the intonational encoding of final boundary tone information between speaking styles.

CORPUS DESCRIPTION

Since we were interested in comparing read and spontaneous speech, we chose a large corpus composed of read and spontaneous utterance pairs, taken from human/machine dialogues, collected in the VOYAGER domain [8]. VOYAGER is a geographical navigation system capable of responding to queries about the location of a set of objects and how to travel between them. Speech was collected in simulation mode, with an experimenter acting as wizard, typing the subject's spontaneous queries to the back end to generate system responses, with the queries being simultaneously recorded. After a 30 minute session, the subject was shown the texts of the utterances he had spontaneously said and asked to read them. These texts exactly matched the spontaneous texts, except that hesitations, filled pauses and false starts were removed. Therefore, each spontaneous utterance in the corpus has a read counterpart said by the same speaker.

Our corpus is composed of 4258 read/spontaneous utterance pairs taken from 89 subjects, 44 males and 45 females, who each said approximately the same number of sentences. To date, all of the utterances have been orthographically transcribed, and about two-thirds have been phonetically transcribed and aligned to their waveforms. In addition, final boundary tones have been labeled in all the utterances.

A study contrasting read and spontaneous speech has already been conducted on a subset of this data. Using utterances from 30 speakers, Soclof and Zue [9] found significant acoustic differences between speech from the two styles. However, most of their analyses were performed at the segmental rather than the suprasegmental level. In addition, they carried out linguistic analyses of this corpus subset, indicating that prosodic markers such as pauses occur in syntactically predictable places, and that such markers are far more common in spontaneous speech than in read. This result supports the assertion that prosodic encoding is more salient in the former style than in the latter.

We computed F_0 every five ms. using an algorithm developed by Secrest and Doddington [10]. This algorithm uses a linear predictive coefficient (LPC) residual signal with a time-varying filter based on the first LPC reflection coefficient. The resulting F_0 contour is smoothed using dynamic programming. A probability of voicing is associated with each F_0 point. For our study, we used only those points with $P(voicing) \geq 0.5$.

GENERAL F_0 COMPARISONS

We performed several general comparisons of F_0 between read and spontaneous speech using the mean, standard deviation and 5^{th} and 95^{th} percentile points to represent the range of our data, computed on a per utterance basis. As Figure 2 reveals, the mean value of F_0 is statistically significantly greater ($p = 0.05$) in spontaneous speech than in read (172.9 Hz *vs.* 167.9 Hz), while their standard deviations are nearly identical (60.8 *vs.* 60.0 Hz). In fact, there is little difference in the F_0 ranges between the two speech styles (95.3 *vs.* 96.2 Hz). Similar trends were observed when the data was partitioned by speaker gender, also shown in Figure 2. The spontaneous and read means of F_0 differed by 6.4 Hz for female speakers and 1.3 Hz for male speakers. We also noted that the range of F_0 are greater in female speech than in male speech (approximately 127 and 63 Hz, respectively, for both speaking styles). This was found to be true not only when the data were compared on a linear scale, but also when compared on a logarithmic scale.

As described in [7], this corpus is composed of utterances of several different sentence types, such as statements and commands. When we partitioned our data along this dimension, we found a number of differences between read and spontaneous F_0. In the interest of brevity, we will report only differences noted in μ_{F_0}. The sentences are categorized into three groups: WH- ques-

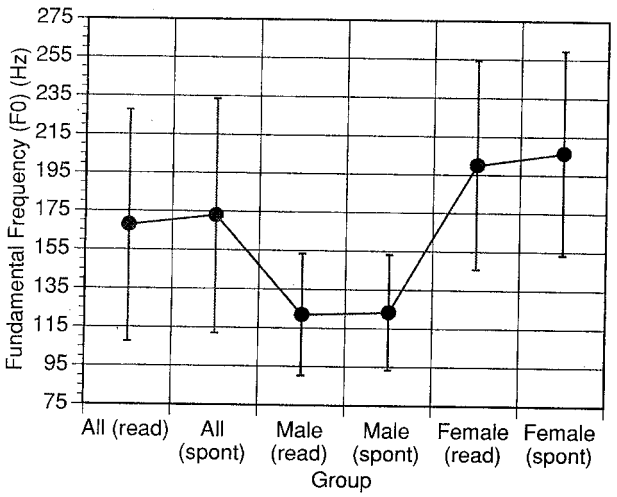

Figure 2: Overall F_0 statistics for the VOYAGER corpus. μ_{F_0} is marked by circles. The error bars extend to $\pm\sigma_{F_0}$. Statistics from all speakers' data are shown in the first two columns, followed by those for the data separated by speaker gender.

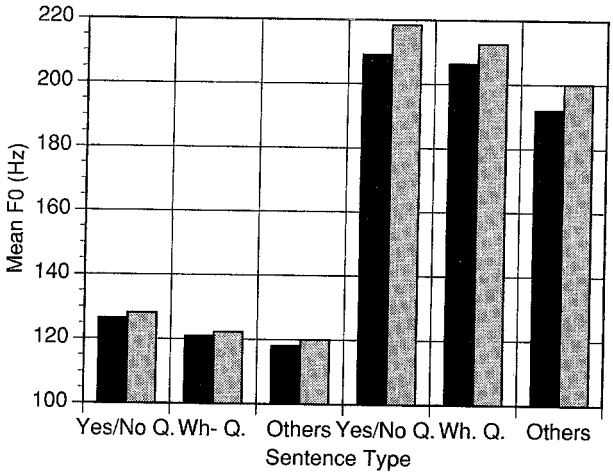

Figure 3: F_0 means for sentence types in the VOYAGER corpus, partioned by speaker gender. Each pair of bars corresponds to read and spontaneous μ_{F_0}, respectively. The first three pairs of bars are for utterances produced by male speakers, and the rest are for those produced by female speakers.

tions, YES-NO questions, which together comprise 90% of the corpus, and all other types, including statements, commands and sentence fragments. These results are shown in Figure 3, with the data separated by speaker gender. The black bars in the figure correspond to read speech and the grey bars to spontaneous speech. For all sentence types, μ_{F_0} is lower for read than for spontaneous speech. We note that the general trends observed previously with respect to speaking style are preserved across all three groups of sentences. In addition, for both speaking styles, YES-NO questions have the highest μ_{F_0} of any sentence type. This can be attributed to the fact that YES-NO questions are more likely to end with high final boundary tones, while all other sentence types tend to end with low final boundary tones. We previously found that for spontaneous speech, 92% of the WH- questions in this corpus ended with low final boundary tones, while 64% of the YES-NO questions ended with high final boundary tones. For read queries, the same percentage of WH-questions had low final boundary tones, but only 58% of the YES-NO questions had high final boundary tones. We suggest that since the read speech in this corpus is non-communicative, there is no need to encode the "yes-no" characteristic of the query as a high final boundary tone for the listener's benefit. This 6% difference in high final boundary tone values between the two speaking styles for YES-NO questions may have contributed to the greater difference in μ_{F_0} for this sentence type, as shown in Figure 3.

FINAL BOUNDARY TONE COMPARISONS

In our previous work [7], we performed a number of experiments analyzing the final boundary tones of the WH- and YES-NO queries in the corpus. Our results indicated that while WH-questions are usually accompanied by low final boundary tones, prosodic encoding of YES-NO questions is more complicated, and depends in part on the speaker's intent. Indirect speech acts, such as *Can you show me how to get there*, are underlyingly polite commands, and as such are more likely to be associated with low final boundary tones than authentic YES-NO questions.

We also performed a pilot study to find possible acoustic correlates of final boundary tones. At that time, we found that these tones occurred during the last syllable of the sentence, and that by measuring F_0 during the last syllable of the utterance, and comparing its mean to the overall mean F_0 for the utterance, we were able to classify and predict final boundary tones with fairly high accuracy (approximately 90%) on a subset of the spontaneous utterances in the VOYAGER corpus.

We have since conducted similar experiments on a larger subset of the corpus on both read and spontaneous speech to compare the salience of final boundary tone encoding between speech from the two styles. Approximately 1200 read and spontaneous utterance pairs were used, taken from queries spoken by 60 different speakers. Two-thirds of the data were used to determine the best measurement for separating high from low boundary tones. The remaining data were used to test the measurement's robustness. In performing these experiments, as before, we tested several acoustic measurements based on F_0 and found that the difference between μ_{F_0} for the final vowel and the rest of the utterance was the best factor for both speaking styles. Evaluations were performed separately for read and spontaneous speech, and for each style, training and testing were performed for male and female speakers both separately and combined.

Figure 4 shows the results of our analysis. For spontaneous speech, final boundary tones were classified with an overall accuracy of about 83%. While these results are not as good as those previously obtained, we note that the data used in this experiment were taken from 60 different speakers, as opposed to 28 in the previous experiment. Since performances on training and test data for this large corpus are quite similar, we conclude that our acoustic measurement generalizes well. For read speech,

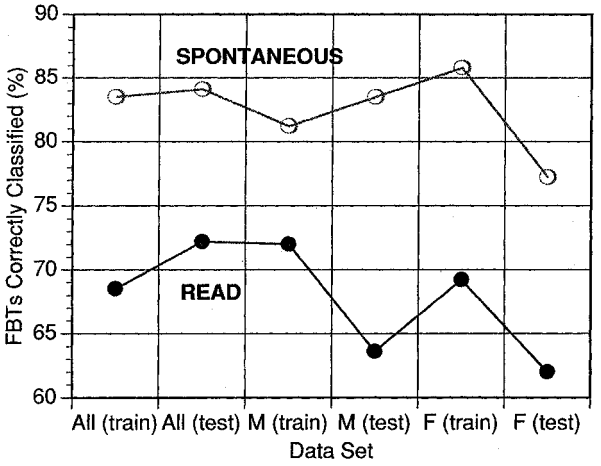

Figure 4: Accuracy of suggested acoustic correlate for classification of final boundary tones (difference between mean F_0 of the last syllable nucleus and mean F_0 over the rest of the utterance).

the overall classification performance is about 68%, a full 15% worse than that for spontaneous speech. Similar trends can be seen when the data are partitioned by gender. These results support our hypothesis that the encoding of final boundary tone information is indeed more salient in spontaneous speech than in read.

DISCUSSION

Based on a number of comparisons between read and spontaneous F_0 when our data is partitioned based on various criteria, we found that the mean of F_0 is greater in spontaneous speech than in read. However, the standard deviation and range of F_0 are virtually identical between the two styles of speech. In addition, we found evidence that prosodic information is encoded in F_0 more clearly in spontaneous speech than in read. Although we believe these results to be compelling, they are not in agreement with those obtained in all other studies involving read and spontaneous speech. For example, in Koopmans-van Beinum's previously-mentioned work [4, 5], the read speech in her corpus had greater mean F_0 and range than the spontaneous speech.

The corpus used in Koopmans-van Beinum's experiments differs from the VOYAGER corpus in several ways, which can account at least partly for the different findings. First, her data consisted of speech from only a single speaker, rather than from a large number of speakers as in VOYAGER. Second, the speech in this corpus was taken from monologues, as opposed to interactive dialogues. And finally, her subject was a professional radio announcer, who we assume was accustomed to reading speech in a *communicative* manner. In contrast, read speech in the VOYAGER corpus was collected by having subjects read from a list without any intent to communicate the meaning of the sentences.

We believe that the communicative intent of the speaker plays an important role in determining his F_0 characteristics. As was previously shown, communicative spontaneous YES-NO queries were more likely to have high final boundary tones indicating their identity than their non-communicative read counterparts. Certainly, it is not sufficient to describe speaking style as merely spontaneous or read. Further examination of the effect of speaker intention on prosodic encoding is warranted.

Another difference between the spontaneous and read speech of our corpus that may have an effect on overall F_0 characteristics for a given style is the presence of filled pauses. For a subset of the VOYAGER corpus taken from 30 speakers, Soclof and Zue found that filled pauses were 12 times more likely to occur in the spontaneous speech than in the read speech. A pilot study of a subset of the filled pauses in the corpus indicates that the mean of F_0 is lowered and its range compressed during filled pauses. However, further study of the intonation of filled pauses is necessary.

In conclusion, we have shown that noticeable differences between read and spontaneous F_0 exist. Furthermore, the encoding of prosodic information in the final boundary tone assignment for syntactic disambiguation is more robust in spontaneous speech than in read. Researchers interested in providing prosodic aids to speech understanding should therefore conduct their studies using speech collected under the paradigm in which understanding is required, since speakers are much more likely to provide proper prosodic markings when their intent is to communicate.

REFERENCES

[1] A. Waibel. "Prosody and Speech Recognition", Ph.D. thesis, Carnegie-Mellon University, 1988.

[2] A. Aull. "Lexical Stress and its Application in Large Vocabulary Speech Recognition", S.M. Thesis, Massachusetts Institute of Technology, 1984.

[3] J. Pierrehumbert. "The Phonology and Phonetics of English Intonation," Ph.D. Thesis, Massachusetts Institute of Technology, 1980.

[4] F. Koopmans-van Beinum. "Spectro-Temporal Reduction and Expansion in Spontaneous Speech and Read Text: The Role of Focus Words," *Proceedings of the 1990 International Conference on Spoken Language Processing*, Kobe, Japan, 1990.

[5] F. Koopmans-van Beinum. "A Peak-And-Level Model for Focus Words in Read and Spontaneous Natural Speech and in Synthetic Speech," *Proceedings of the Second European Conference on Speech Communication and Technology*, Genova, Italy, 1991.

[6] P. Howell and K. Kadi-Hanifi. "Comparison of prosodic properties between read and spontaneous speech material," from *Speech Communication 10.* Amsterdam: North-Holland Publishing Company, 1991.

[7] N. Daly and V. Zue. "Acoustic, Perceptual, and Linguistic Analyses of Intonation Contours in Human/Machine Dialogues," *Proceedings of the 1990 International Conference on Spoken Language Processing*, Kobe, Japan, 1990.

[8] V. Zue, J. Glass, D. Goodine, H. Leung, M. Phillips, J. Polifroni, and S. Seneff. "The VOYAGER Speech Understanding System: Preliminary Development and Evaluation," *Proceedings, International Conference on Acoustics, Speech and Signal Processing*, Albuquerque, NM, USA, 1990.

[9] M. Soclof and V. Zue. "Collection and Analysis of Spontaneous and Read Corpora for Spoken Language System Development," *Proceedings of the 1990 International Conference on Spoken Language Processing*, Kobe, Japan, 1990.

[10] B. Secrest and G. Doddington. "An Integrated Pitch Tracking Algorithm for Speech Systems," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Boston, MA, USA, 1983.