A prosodically labeled database of spontaneous speech

*M. Ostendorf†, I. Shafran†, S. Shattuck-Hufnagel‡, L. Carmichael†, and W. Byrne**

†University of Washington, Seattle, WA
‡MIT, Boston, MA * JHU, Baltimore, MD

Abstract

This paper describes a prosodically labeled database of conversational speech, representing a subset of the Switchboard and Callhome corpora. The prosodic transcription system is a simplification of the ToBI system aimed at phenomena that would be most useful for automatic transcription and linguistic analysis of conversational speech. The transcription method and a distributional analysis of the types of prosodic events are described.

1. Introduction

The usefulness of speech databases for statistical analysis is substantially increased when the database is labeled for a range of linguistic phenomena, providing the opportunity to improve our understanding of the factors governing systematic suprasegmental and segmental variation in word forms. One of the most promising domains for such analysis is the relationship between prosodic structure (i.e. the structures governing utterance-level intonation, timing, amplitude etc.) and surface phonetic variation. Prosodic labels and phonetic alignments are available for some corpora of connected communicative speech (e.g. the BU FM Radio News Corpus of professionally read broadcast news [1]), but only a few limited samples of spontaneous conversational speech have been prosodically labeled.

To fill this gap, substantial samples of the Switchboard corpus of spontaneous telephone-quality dialogs were labeled using a subset of the ToBI system for transcribing pitch accents, boundary tones and prosodic constituent structure [2, 3]. Samples were selected to include longer portions of a number of dialogs, totaling roughly 4.7 hours of speech. Details of the corpus and the transcription conventions are provided here, followed by an overview of the distributional characteristics of the corpus.

2. Corpus Description

The corpus is based on about 4.7 hours of hand-labeled conversational speech (excluding silence and noise) from 63 conversations of the Switchboard corpus [4] and 1 conversation from the CallHome corpus. There are 45 conversation sides from male speakers and 83 from female speakers, and about 2/3 of the labeled data is from females. (Some speakers talked more than others, and some conversation sides were not completely labeled.) We chose to focus on longer stretches of conversation rather than getting large numbers of speakers, because of our interest in relating prosody and discourse structure and because of the need to make the labelling task more time efficient. The corpus contains about 67k word instances (excluding silences and noise), and within this set are roughly 1.5k filled pauses.

The data is orthographically transcribed by hand and corrected with multiple passes of human review, so that the orthographic transcription error is very small. These transcriptions are time-aligned to the waveform using segmentations available from Mississippi State University (http://www.isip.msstate.edu/projects/switchboard/index.html) and a large vocabulary speech recognition system developed at the JHU Center for Language and Speech Processing for the 1997 Language Engineering Summer Workshop (www.clsp.jhu.edu/ws97) . The system consisted of 12 Gaussian mixture, tied-state, cross-word triphone acoustic models [5] trained using the HTK Hidden Markov Model Toolkit (http://htk.eng.cam.ac.uk/). The acoustic features used were PLP-Cepstra, and the pronunciation lexicon used for training was derived from Pronlex (http://www.ldc.upenn.edu/Catalog/LDC97L20.html).

All conversations have been analyzed using a high quality pitch tracker [6] to obtain fundamental (F0) frequency contours. Because the pitch tracker picks up the pitch of the background speaker in regions where there is crosstalk (even for very low levels of crosstalk), it was necessary to post-process the data and eliminate F0 values during regions of crosstalk determined automatically from the time alignments.

An advantage of working with the Switchboard corpus is that it has been hand annotated with a variety of linguistic structures, including part-of-speech tags and syntactic parses as part of the Penn Treebank (http://www.cis.upenn.edu/ treebank/), as well as disfluencies, discourse markers and dialog act labels [7, 8].

3. Prosodic Labelling

The prosody transcription system is a variant of the ToBI labelling system, which is based on the intonational theory of Beckman and Pierrehumbert [9] and the break index labelling system of Price et al. [10]. It permits notation of the prominent syllables and structural boundaries in digitized spoken utterances in a simplified manner. It also provides a way to deal with the disfluencies that are common in spontaneous conversational speech, and to indicate labeller uncertainty about a particular transcription.

The elements of this prosodic transcription alphabet include **breaks**, which indicate the depth of the boundary after each word, and **tones**, which indicate syllable prominence and tonal boundary markers. A script using xwaves tools displays the waveform files, estimated F0 tracks and label files for either one or both of the two speakers in the conversation. The label files include word, tone, break and miscellaneous tiers. Labellers use a tier-specific menu to select the appropriate labels for each prominent syllable and word boundary.

Break indices labelled after every word included:

0: well-formed phonetic modification across the boundary, such as 'doncha' for 'don't you', or other cues to particularly tight junctures;

1: normal word boundary, provided as default at every word boundary in the initial label file;

2: a larger boundary than normal, often cued by some duration lengthening but not enough to indicate a planning difficulty or an intonational phrase boundary;

3: the boundary of an intermediate intonational phrase, cued by some pre-boundary lengthening and a phrase tone (see below); and

4: the boundary of a full intonational phrase, cued by more pre-boundary lengthening and a full boundary tone, i.e. a combination of two tones (again, see below). When a sequence of intermediate phrases seems to group together, it forms a single full intonational phrase.

As in ToBI, each 3 and 4 must be labelled with a final boundary-related tone, and must contain at least one pitch accent. The diacritic "-" was used for indicating uncertainty in labelling break labels, e.g. a "4-" would indicate uncertainty as to whether the break constituted a full vs. intermediate intonational phrase. Breaks were not assigned at non-words such as the transcribed [PAU], [laughter], etc. In addition, filled pauses (um, uh, er, etc.) and back-channels (yeah, uh-huh) were usually marked with break "X", since these cases proved difficult for the labellers and we felt that they should be studied separately before deciding on a labelling convention. In addition, "you know" sometimes fell in this class, since it sometimes served as a filled pause. Labellers were allowed to mark tone/break structure for clear cases.

A "p" diacritic was used for indicating a disfluent word boundary in a prosodic phrase where the phrase is eventually completed. A "p" diacritic was only used in combination with breaks 1 and 2: a 1p corresponded to a word fragment or a word that was perceived to be cut short, and a 2p was used to indicate hesitation-related lengthening, a word boundary before a breath that does not have the tonal cues associated with a 3 or 4, or incomplete prosodic phrases. Initially, the data was labeled with additional diacritics to separately indicate an incomplete prosodic phrase and other phenomena, but these were not labeled reliably, so they were omitted from the final label set.

Tones were labelled at phrase boundaries, as follows. At each 4 or 4- break (full intonational phrase), the possible tones included:

L-L% final fall to bottom of range, often but not necessarily utterance final

L-H% fall-rise (rise can sometimes be small or flat), sounds like more to come, often called a continuation rise

H-L% slight fall but not to bottom of range, sounds more incomplete than final fall

H-H% rise to high point in range, often called a question rise since it us used in Y-N questions. Used in a statement, it gives the impression of doubt or request for agreement.

At each 3 or 3- Break (intermediate intonational phrase), the possible tones included:

L- ends in a low value relative to the rest of the phrase

H- ends in a high value relative to the last pitch accent

!H- ends at a mid-range point, small fall if last pitch accent is high

In a departure from the standard ToBI framework, prominent syllables were marked only with "*" for indicating a pitch accent (tonally cued prominence) or "*?" for a possible prominence (i.e. uncertainty about presence of a pitch accent). The specific tone markers were omitted to reduce the labelling time. Another marker for a non-tonally-cued prominence was initially made available to the labellers, but since its use was not consistent, it was not included in the final inventory.

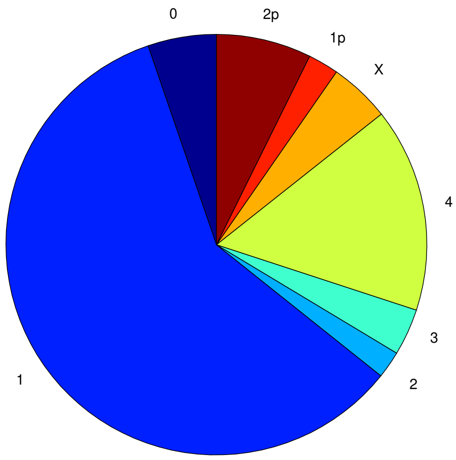

Figure 1: ToBI break indices in the corpus.

Labellers were encouraged to attend to both perception and the visual display of the F0 contour, keeping an eye out for pitch halving/doubling and for predictable segmental effects on F0. They were discouraged from changing the word time marks unless the existing marks would result in a misalignment of a prosodic symbol, i.e. placed a boundary tone or pitch accent in the wrong word or outside of its word. They were allowed to correct orthographic transcription errors, but these were rare.

Initial labelling was carried out by 5 undergraduates at MIT who had no prior training in linguistics, speech science or prosodic labelling. They received 6 hours of instruction from experienced ToBI labellers, including practice labelling and discussion of an already-labelled "reference" dialog, and were equipped with an 8-page training manual. Several of the labellers were not native speakers of American English.

An initial consistency study was run on one conversation labeled by all the students and one more experienced labeller. The study revealed problems with some of the new diacritics, so these were dropped, as mentioned above. Unfortunately, the consistency of the labellers was still not sufficiently high, so we decided to have a single more experienced labeller (a graduate student in linguistics trained in the ToBI labelling system) retranscribe the entire corpus.

4. Distributional Analyses

The distribution of various break indices in the corpus is shown in Fig. 1. About one in ten words followed a disfluent word boundary (1p or 2p), but in fact the percentage is somewhat higher since this excludes most breaks after filled pauses. Of the 14.5K cases where there was a pause after the word, 31% corresponded to a disfluent word boundary.

Of the complete words in the corpus (excluding filled pauses and word fragments), about one third were found to have a pitch accent. Very few words had more than one accent (less than 0.2%).

5. Discussion

In summary, we have described a prosodically labeled corpus of conversational speech and preliminary distributional analyses of this data. Further analyses of the data is needed, including a self-consistency check of the labeller used in the second pass of transcription. In addition, we hope to enrich this data by associating the time alignments with syntactic and dialog act labels that are also available for this corpus but were based on a slightly different version of the orthographic transcriptions.

Recent findings support the claim that hierarchically-organized prosodic constituent structure and prominence patterns influence phonetic implementation phenomena such as glottalization of word-onset vowels, constituent-final acoustic lengthening, constituent-initial articulatory strengthening etc. We envision that the availability of a substantial prosodically labeled database of spontaneous speech, which includes the normal range of variation in surface phonetic form, will provide the resources required to study and understand the relationship between prosodic structure and phonetic variation. A preliminary analysis is reported in [11]. In addition, the tone labels at phrase boundaries will make it possible to conduct an empirical study of the relation between tones and linguistic structures such as syntactic phrase type and dialog acts.

Acknowledgments

This work was supported in part by NSF, award numbers IIS-9618926, BCS-98-20126, and IIS-0095940, and by NIH, award number DC02125. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

6. References

[1] Ostendorf, M., Price, P. and Shattuck-Hufnagel, S. "The Boston University Radio News Corpus," Boston University Technical Report ECS-95-001, 1995.

[2] Silverman, K., et. al., "ToBI: A Standard for Labeling English Prosody," In *Proceedings of ICSLP*, 867–870, 1992.

[3] Pitrelli, J., Beckman, M. and Hirschberg, J. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of the International Conference on Spoken Language Processing*, Yokohama, Japan, **1**, 123-126, 1994.

[4] Godfrey, J.J., Holliman, E.C., and McDaniel, J., "SWITCHBOARD: Telephone Speech Corpus for Research and Development," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* pp. I-517-520, 1992.

[5] Byrne W., Finke M., Khudanpur S., McDonough J., Nock H., Riley M., Saraclar M., Wooters C. and Zavaliagkos .G, "Pronunciation Modelling Using a Hand-Labelled Corpus for Conversational Speech Recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing,* Seattle, WA, 1998.

[6] Talkin, D., "Pitch Tracking," in *Speech Coding and Synthesis,* ed. W. B. Kleijn and K. K. Paliwal, Elsevier Science B.V., 1995.

[7] Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Meteer, M., and Van Ess-Dykema, C., "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech," *Computational Linguistics,* 26:3, 2000.

[8] Shriberg, E., Bates, R., Taylor, P., Stolcke, A., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., and Van Ess-Dykema, C., "Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?," *Language and Speech,* 41:3-4, 439-487, 1998.

[9] Beckman, M. and Pierrehumbert, J., Intonational structure in Japanese and English. *Phonology Yearbook*, **3**, 255–309, 1986.

[10] 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**, 2956-2970, 1991.

[11] Shafran, I. Ostendorf, M., and Wright, R., "Prosody and phonetic variability: Lessons learned from acoustic model clustering," this proceedings, 2000.