Towards Speaker Independent Features for Information Extraction from Meeting Audio Data

Sarah Simpson and Yoshihiko Gotoh

Speech and Hearing Group, Department of Computer Science,
University of Sheffield, UK
s.simpson@dcs.shef.ac.uk, y.gotoh@dcs.shef.ac.uk

1 Extended Abstract

The aim of early information extraction systems involved sentence segmentation, topic segmentation and named entity extraction from text using typographic cues, such as punctuation, to define the structure of the passage. With developments in speech recognition technology the research progressed to apply these tasks to spoken language. Initially these systems were similar to the text-based system, based purely on lexical information (for example [1]). Not only is this a rather simplistic viewpoint, it also disregards the many cues available in the waveform. One set of acoustic cues that has been successfully introduced to access the structure of spontaneous speech is prosodic information (for example [2]). Generally, word transcripts are aligned with the acoustic waveforms to obtain word boundaries, after which, features based on pitch, energy and duration are estimated from the waveform. Typical features include precedent pause, subsequent pause, mean energy, pitch range, pitch onset and pitch offset.

A major challenge in systems that incorporate prosodic information is to find a set of features that is speaker independent. Absolute values of pitch and energy vary greatly with speaker and place so, to counteract this, such algorithms employ various normalization techniques. However, more research is required to fully assess how much these features vary among speakers, especially in different conditions. The aim of this study is to find simple acoustic features that are speaker independent within meeting conditions.

For the purpose of this study ten randomly selected meetings from the ICSI meetings corpus were used. For each meeting audio file, the fundamental frequency was estimated using the YIN algorithm [3] and the energy envelope was calculated. These were then used to compute the following selected features: energy change across pauses, energy change across the acoustic units, pitch change across pauses and energy difference across units. It should also be noticed that no absolute pitch or energy values were used. As well as being intrinsically more speaker independent than absolute values, the chosen features represent the amount of ‘change’ in the signal. Given the emphasis that perceptual systems place on change in the environment it seems intuitive that speakers use this, either consciously or subconsciously, to attract the listeners’ attention to the key elements of the dialogue. Another point of note is that, in contrast to other studies that use word boundaries defined directly by the speech recognition transcript, the temporal boundaries of acoustic events here estimated using an

algorithm based on [4]. The reason for this is that the lexical concept of a word does not necessarily have a one to one relationship with spoken units. Therefore while slight computational errors may be introduced it removes the assumption that spoken language has the same grammatical units as in the written form. These boundaries were also used to compute the fifth feature, pause duration.

The first stage in the investigation was to observe the distribution of each feature, across the different speakers, by plotting histograms. For all the features, there appeared to be a good correlation amongst speakers. A curve was then fitted to the collated data for each feature. The accuracy of each model was examined by comparing the predicted distribution to the observed results of a further five meetings (see Fig 1).

Once the features that display good speaker independence have been identified their models can be combined with a priori knowledge regarding how speakers use prosody to indicate key information. The ultimate goal is to produce a scoring system that highlights the units in the acoustic signal that contain the most information.

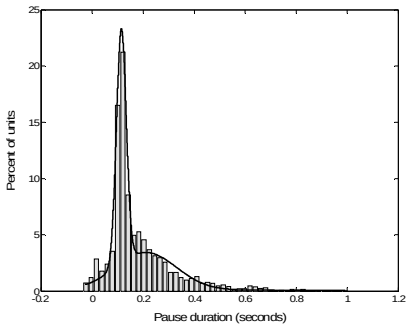

Fig. 1. Comparison between observed distribution of pause durations in a test meeting (gray bars) and the model fitted to the training data (black line)

2 Acknowledgement

This work was supported by EPSRC grant GR/R42405.

References

1. Kubala, F., Schwartz, R., Stone, R. and Weischedel, R.: Named entity extraction from speech. Proc. DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne VA (1998).
2. Koumpis, K. and Renals, S.: Automatic summarization of voicemail messages using lexical and prosodic features. ACM Transactions on Speech and Language Processing (2004) Submitted for publication.
3. de Cheveigne, A. and Kawahara, H.: YIN, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am. **111** (2001) 1917-1930.
4. Salomon, A., Espy-Wilson, C.Y., and Deshmukh, O.: Detection of speech landmarks: Use of temporal information. J. Acoust. Soc. Am. **115** (2004) 1296-1305.