Towards Speaker Independent Features for Information Extraction from Meeting Audio Data

Sarah Simpson and Yoshihiko Gotoh

Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, UK

E-mail: s.simpson@dcs.shef.ac.uk, y.gotoh@dcs.shef.ac.uk

1. Introduction

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 and consequently, disregarding the many cues available in the waveform.

One set of acoustic cues that has been successfully introduced is prosodic information.

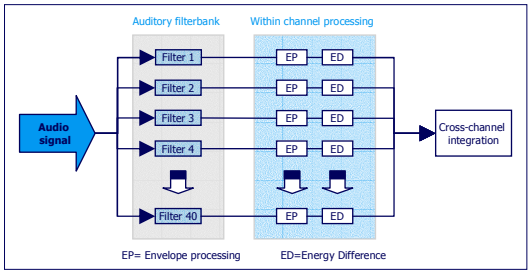
A major challenge in systems that incorporate prosodic information is to find a set of features that is speaker independent. Absolute values of pitch and energy vary greatly with speaker and place so, to counteract this, such algorithms employ various normalization techniques.

The aim of this study is to find simple acoustic features that are speaker independent within meeting conditions.

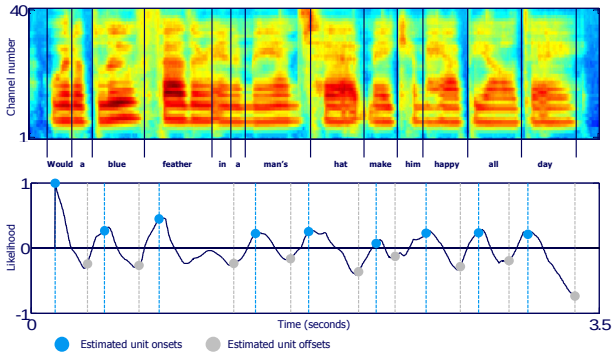
2. Unit boundaries

It is unlikely that the lexical concept of a word does not have a one to one relationship with spoken units.

Word boundaries were not defined directly by the speech recognition transcript. Instead unit boundaries were estimated directly from the audio signal.



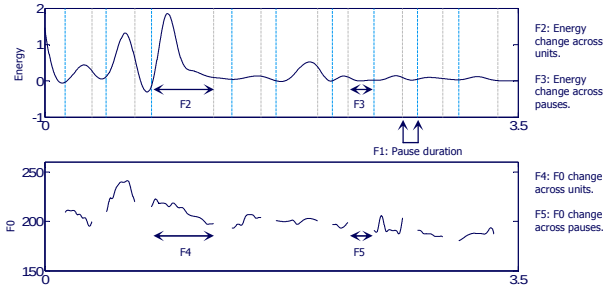
Algorithm adapted from [2].



3. Features

Features were calculated using:

- Unit boundaries;
- Energy envelope (from Envelope processing stage above);
- Fundamental frequency (estimated using the YIN algorithm [1]).



F2-F4 are calculated by estimating the gradient of the values with the specified unit boundaries.

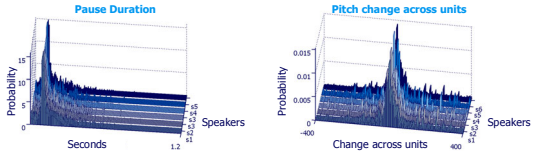
No absolute pitch or energy values were used.

The chosen features are more speaker independent and represent the amount of 'change' in the signal.

4. Feature models

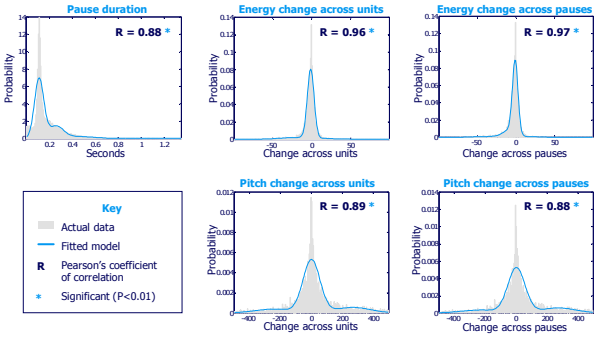
The first stage in the investigation was to observe the distribution of each feature, across the different speakers, by plotting histograms.

For all the features, there appeared to be a good correlation amongst speakers.



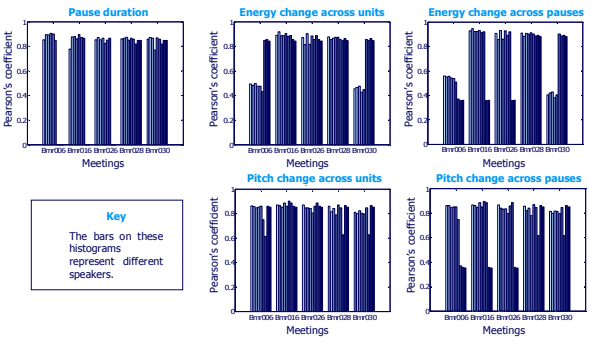
The data for each feature was collated and a histogram plotted. The area was normalized to 1 to give an approximate PDF.

A gaussian mixture was then fitted.



5. Results

The accuracy of each model was examined by comparing the predicted distribution to the observed results of a further five meetings.



6. Summary

All features models had a significant (P<0.01) correlation with the test meetings.

Pause duration and the pitch change across the units appears to be the most consistent.

This indicates that these simple acoustic features have a good degree of speaker independence.

However, this result needs to be checked with meetings from different corpora.

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.

For example: A person increases the volume (amount of energy) of their voice to draw attention to the current unit or key units that directly follow. Therefore a large increase in energy across a unit would result a high score for the current unit and the following unit. The units with the highest scores would be deemed to be the most important.

The ultimate goal is to produce a scoring system that highlights the units in the acoustic signal that contain the most information.

7. References

1. de Cheveigne, A. and Kawahara, H.: YIN, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am. **111** (2001) 1917-1930.
2. 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.

8. Acknowledgement

This work was supported by EPSRC grant GR/R42405.