Compensation for the effects of reverberation on automatic speech recognition: a perceptually-inspired approach based on weighting of parallel acoustic models

Guy J. Brown¹, Kalle J. Palomäki² and Amy V. Beeston¹
¹Department of Computer Science, University of Sheffield, UK
²Department of Computer and Information Science, Aalto University School of Science and Technology, Finland
g.brown@dcs.shef.ac.uk, a.beeston@dcs.shef.ac.uk, kpalomak@cis.hut.fi

Background
Watkins (2005) has shown that listeners use information about the preceding context of a reverberated test word to help them identify it. This suggests a mechanism of perceptual constancy that confers robustness in reverberant environments. Watkins’ experiments focused on one particular speech identification task (“sir” or “stir”), and used a synthesised continuum to measure the “sir/stir” category boundary. Beeston et al (2010) extended Watkins findings using natural speech and a wider range of consonants (/p/, /t/, /n/).
Here we focus on the development of a computer model, which aims to replicate the pattern of consonant confusions observed in Beeston et al’s data.

Aims of the current study
To implement a computer model of perceptual compensation for reverberation based on acoustic model selection.
To determine whether the computer model is able to match the pattern of confusions evident in human data.
To compare the performance of a fully autonomous model with one in which ‘oracle’ information is given about the appropriate acoustic model to use.

Perceptual experiment
Test material was drawn from the Articulation Index (AI) corpus (Wright, 2005), 80 utterances of the form CW1 CW2 TEST CW3

Accessibility of the context words and test words varied independently, as described by Watkins (2005).
The reverberation of the context words and test words was varied individually, as described by Watkins (2005).

A perceptual compensation effect is observed: confusions with a ‘far’ test word and ‘near’ context are reduced if the context is also reverberated at the ‘far’ distance.

Conceptual model
Perceptual compensation for the effects of reverberation could be viewed as an acoustic model selection process.
Analysis of the speech preceding a test word informs selection of an appropriate acoustic model.

Perception is optimal when the reverberation conditions of the context speech and test word are the same.
When the reverberation applied to the context speech and target word differs, a mismatch occurs and consonant confusions increase.

Analysis of confusions
Pearson’s phi-squared statistic used to determine similarity of human and model confusions (Jürgens & Brand, 2009).
Each row of human and model confusion matrices compared as 2x4 contingency table. For identical distributions $\phi^2 = 0$, for non-overlapping distributions $\phi^2 = 1$.

Oracle feature stream selection
The model reproduces the main confusions evident in the human data; $\phi^2 \leq 0.1$ in all but one condition.

Computer model
The simulation is based on a Hidden Markov Model (HMM) automatic speech recognition system.
40 monophone models and a silence model. Initial training on TIMIT corpus, then adaptation on the subset of the AI corpus used by Beeston et al.
Acoustic features were 12 mel-frequency cepstral coefficients (MFCCs) + deltas + accelerations.

The recogniser was trained with feature vectors consisting of two blocks of 36 acoustic features, obtained from speech filtered with the ‘near’ and ‘far’ room impulse responses.
The HMMs for the combined features were then split after training to give separate ‘near’ and ‘far’ acoustic models.

Combining feature streams
During decoding, for each feature vector $x(t)$ at time $t$, the observation state likelihoods are computed from the HMMs for both feature streams.
We use $p(x(t)|\lambda_n)$ and $p(x(t)|\lambda_f)$ to denote the likelihood computed from the ‘near’ and ‘far’ acoustic models respectively.
The combined near-far observation state likelihood is a weighted sum of likelihoods in the log domain:

$$\log \{ p(x(t)|\lambda_n) \} = \alpha(t) \log \{ p(x(t)|\lambda_n) \} + (1-\alpha(t)) \log \{ p(x(t)|\lambda_f) \}$$

The weighting factor $\alpha(t)$ is adjusted dynamically according to the acoustic conditions, $\alpha(t)$→0 if reverberant and $\alpha(t)$→1 if dry.

Determining the weighting factor
Simplest approach: use an ‘oracle’ value of $\alpha(t)$, assuming that context reverberation condition is known.
Fully autonomous model: estimate the value of $\alpha(t)$ from the context speech.
Here, we use the mean-to-peak ratio (MPR) of the context speech envelope as a measure of the amount of reverberation present.
A Gaussian classifier detects a ‘near’ or ‘far’ context using the MPR as input (83% correct classification on test set).

Conclusions
The model gives a good match to the pattern of confusions in the human perceptual compensation data.
The ‘oracle’ and fully-autonomous models give similar confusion patterns, although the overall word recognition rate is lower for the latter.

Acknowledgments
GB and AVB were supported by EPSRC. KJP was supported by the Academy of Finland. Thanks to Tim Jürgens for assistance with the phi-squared metric.

References