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

Guy J. Brown[1], Kalle J. Palomäki[2] and Amy V. Beeston[1]

[1]Department of Computer Science, University of Sheffield, UK
[2]Department 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

The University Of Sheffield.

Aalto University School of Science

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/, /k/).
- 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

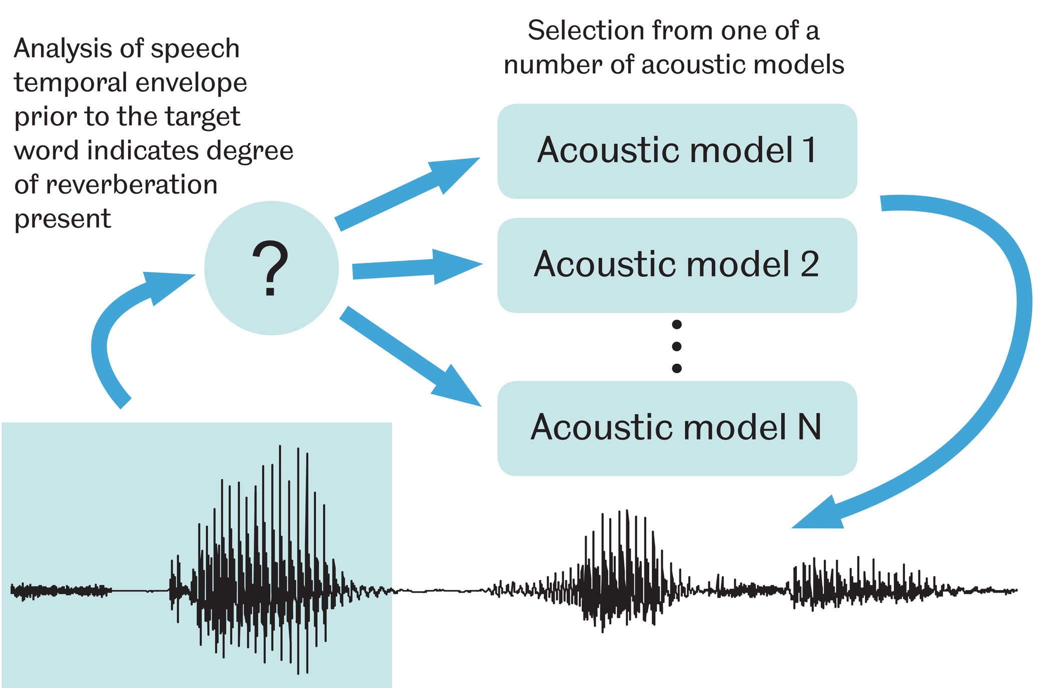
- Context words (CW) were drawn from a limited set and the test word was SIR, SKUR, SPUR or STIR.
- The reverberation of the context words and test words was varied independently, as described by Watkins (2005).
- The reverberation was varied according to the source-receiver distance in an L-shaped conference room (impulse responses recorded by Watkins).

		Test word distance	
		0.32m	10m
Context distance	0.32m	near-near	near-far
	10m	far-near	far-far

- 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

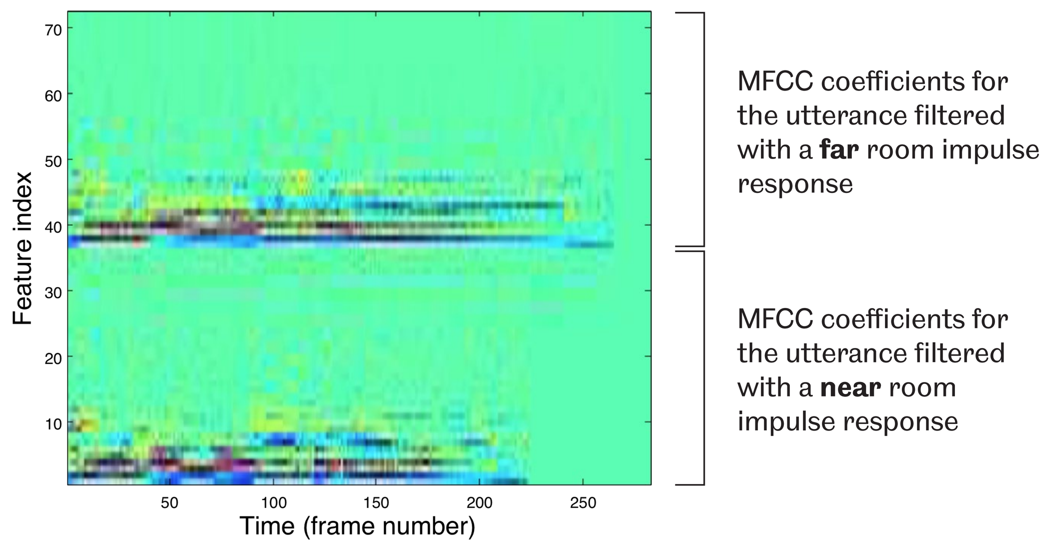
- Perceptual compensation for the effects of reverberation could be viewed as an **acoustic model selection** process.
- Analysis of the speech preceding a test word informs selection of an appropriate acoustic model.



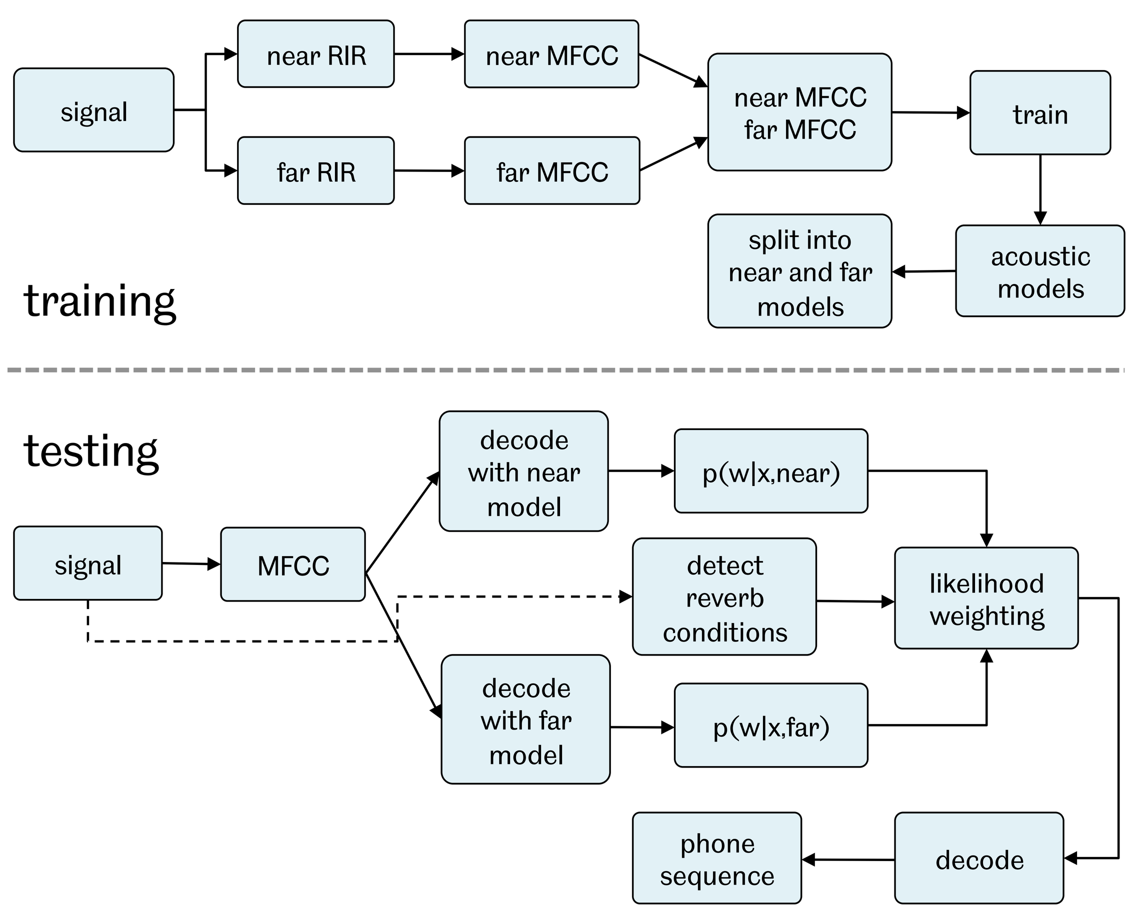
- Performance is optimal when the reverberation conditions of the context speech and test word are the same.
- When the reverberation applied to the context speech and target word differs, a mismatch occurs and consonant confusions increase.

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:

$$\log[\,p(x(t)|\,\lambda_{n,f}] = \alpha(t)\log[\,p(x(t)|\,\lambda_n)\,] + (1-\alpha(t))\log[\,p(x(t)|\,\lambda_f)\,]$$

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

Analysis of confusions

- Pearson's phi-squared statistic used to determine similarity of human and model confusions (Jurgens & 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.

Human near-near					Oracle model near-near					ϕ^2
	SIR	SKIR	SPIR	STIR		SIR	SKIR	SPIR	STIR	
SIR	19	0	0	1	SIR	16	0	0	4	0.0514
SKIR	0	20	0	0	SKIR	0	19	0	1	0.0256
SPIR	0	1	18	1	SPIR	1	0	19	0	0.0757
STIR	0	0	0	20	STIR	0	1	0	19	0.0256

Human near-far					Oracle model near-far					ϕ^2
	SIR	SKIR	SPIR	STIR		SIR	SKIR	SPIR	STIR	
SIR	18	0	0	2	SIR	18	1	1	0	0.1000
SKIR	3	15	0	2	SKIR	3	17	0	0	0.0531
SPIR	7	2	10	1	SPIR	3	1	15	1	0.0733
STIR	8	1	1	10	STIR	9	3	0	8	0.0570

Human far-far					Oracle model far-far					ϕ^2
	SIR	SKIR	SPIR	STIR		SIR	SKIR	SPIR	STIR	
SIR	16	1	1	2	SIR	11	2	2	5	0.0720
SKIR	0	16	0	4	SKIR	1	18	0	1	0.0729
SPIR	2	1	14	3	SPIR	2	0	18	0	0.1125
STIR	1	0	0	19	STIR	0	0	0	20	0.0256

- Few confusions in the near-near condition. In the near-far condition, the predominant confusion is STIR → SIR.
- The STIR → SIR confusion is resolved in the far-far condition in both the human and model confusion matrices.

Feature stream selection based on MPR of envelope

- Again, predominant human confusions are well-reproduced by the model, but overall recognition rate is lower.

Human near-near					MPR model near-near					ϕ^2
	SIR	SKIR	SPIR	STIR		SIR	SKIR	SPIR	STIR	
SIR	19	0	0	1	SIR	16	0	0	4	0.0514
SKIR	0	20	0	0	SKIR	0	19	0	1	0.0256
SPIR	0	1	18	1	SPIR	1	0	17	2	0.0590
STIR	0	0	0	20	STIR	1	1	1	17	0.0811

Human near-far					MPR model near-far					ϕ^2
	SIR	SKIR	SPIR	STIR		SIR	SKIR	SPIR	STIR	
SIR	18	0	0	2	SIR	18	0	1	1	0.0333
SKIR	3	15	0	2	SKIR	3	17	0	0	0.0531
SPIR	7	2	10	1	SPIR	5	1	14	0	0.0583
STIR	8	1	1	10	STIR	8	3	0	9	0.0513

Human far-far					MPR model far-far					ϕ^2
	SIR	SKIR	SPIR	STIR		SIR	SKIR	SPIR	STIR	
SIR	16	1	1	2	SIR	14	1	2	3	0.0167
SKIR	0	16	0	4	SKIR	2	16	0	2	0.0667
SPIR	2	1	14	3	SPIR	3	0	16	1	0.0583
STIR	1	0	0	19	STIR	0	0	0	20	0.0256

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

GJB 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

Beeston, A.V., Brown, G. J., Watkins, A. J. & Makin, S. J. 2010. Perceptual compensation for reverberation: human identification of stop consonants in reverberated speech contexts. *British Society of Audiology Annual Conference*, University of Manchester, September 8th-10th.

Jürgens, T. & Brand, T. 2009. Microscopic prediction of speech recognition for listener with normal hearing in noise using an auditory model. *J Acoust Soc Am*, 125(5), 2635–2648.

Watkins, A. J. 2005. Perceptual compensation for effects of reverberation in speech identification. *J Acoust Soc Am*, 118(1), 249-262.

Wright J. 2005. *Articulation Index*. Linguistic Data Consortium, Philadelphia.