A computer model of perceptual compensation for reverberation: evaluation on a consonant identification task

Guy J. Brown and Amy V. Beeston
Department of Computer Science, University of Sheffield
{g.brown,a.beeston}@dcs.shef.ac.uk

Introduction

- 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.
- Here we address the following research questions:
  - Is perceptual compensation for the effects of reverberation also apparent in a more naturalistic consonant discrimination task (/p/, /t/, /k/)?
  - How does the reverberation-robustness of a conventional automatic speech recognition (ASR) system compare with human listeners?
  - Does an auditory model with an efferent processing circuit effect compensation for reverberation in a similar manner to human listeners?
- Our eventual aim is to build a human-like 'constancy front-end' for ASR.

Test Material

- 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.
- All utterances were low-pass filtered to 4 kHz to avoid ceiling effect when testing for consonant confusions.
- Perceptual constancy was investigated by varying reverberation of the context words and test words 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

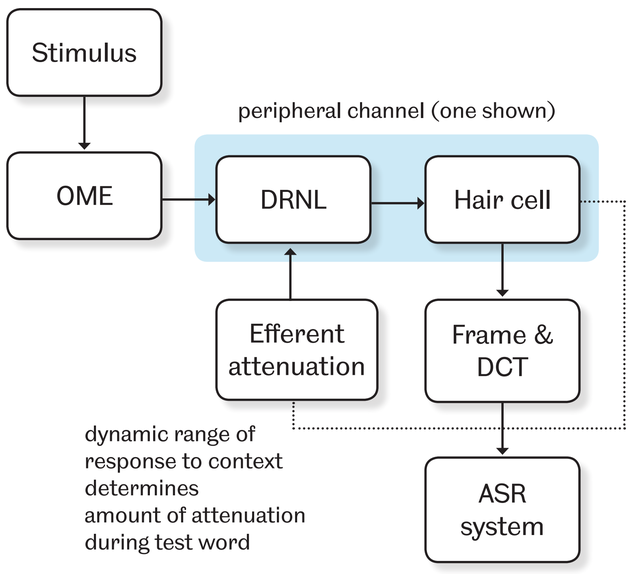
- After low-pass filtering and convolution with the room impulse response, a filter was applied to correct for the response of the headphones used in listening tests.
- Detailed perceptual studies are reported in a companion poster.

Speech Recogniser

- A speech recogniser was developed using the hidden Markov model toolkit (http://htk.eng.cam.ac.uk/).
- Phone-level (rather than word-level) recognition was required in order to assess consonant confusions.
- 39 monophone models were trained, with observations modelled with 20 Gaussian mixtures per state.
- In the AI corpus, phonetic transcripts are only provided for the target words. The context words were expanded to a phone sequence using the CMU pronunciation dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict).
- The recogniser was initially trained on the TIMIT corpus (which is provided with detailed phonetic transcriptions) and then further embedded training was performed on the AI corpus.
- A baseline ASR system was trained using 12 MFCC features or 13 DCT-transformed auditory features, plus deltas and accelerations.
- Semi-forced alignment was used; the recogniser was told the identity of the context words and was required to identify the test word only.

Auditory Model

- The auditory model is a modification of the Ferry & Meddis (2007) model of auditory efferent processing.



- Efferent activity is modelled as an attenuation in the nonlinear path of a dual-resonance nonlinear filterbank (DRNL).
- The amount of efferent attenuation is determined by measuring the dynamic range of the preceding speech context.
- The model has previously been shown to give a good match to listener data in Watkins' (2005) 'sir'/'stir' identification task (Beeston & Brown, 2010).

Evaluation

- Human and machine performance were compared in terms of percentage error and relative information transmitted (RIT).
- RIT is an information-theoretic metric that reflects the distribution of errors in the confusion matrix.
- The subject (human or ASR system) is regarded as a channel that accepts input and produces output, and RIT measures its information transfer characteristics:

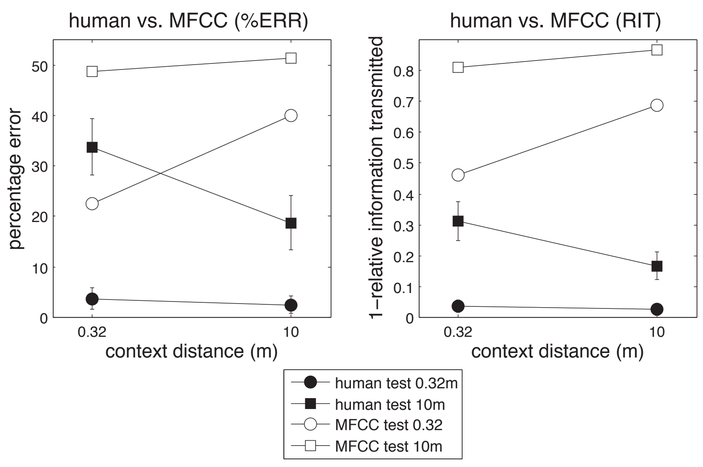
$$RIT = H(X:Y)/H(X)$$

- H(X:Y) is the average mutual information of the input X and output Y, and H(X) is the average self-information (entropy) of the input.

Results

Experiment 1: Comparison of human performance and baseline ASR system

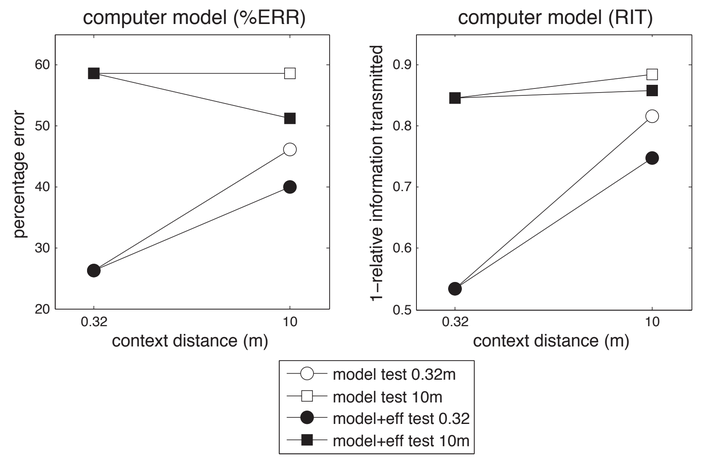
- Human listeners show perceptual compensation; for a 'far' test word (10m) percentage error is high with a 'near' context but lower with a 'far' context.
- This pattern is also observed in the RIT metric (i.e, compensation is apparent as an improvement in the pattern of confusions made by listeners).



- As expected the baseline ASR system has a higher overall error rate than human listeners and does not show compensation.
- For the ASR system, errors are directly related to the amount of reverberation in the test word (error in near-near < far-near < near-far < far-far).

Experiment 2: Auditory model performance with and without efferent circuit

- When the efferent circuit is not engaged, the auditory model behaves similarly to the baseline MFCC system.
- Percentage error is slightly higher, most likely due to nonlinear (level-dependent) behaviour of the DRNL.



- When the efferent circuit is engaged, 'far' context conditions cause 4dB attenuation in the test word.
- This leads to a small amount of compensation, measured as reduced percentage error in the far-far condition compared to the near-far condition.
- However, compensation is not apparent when measured in terms of RIT.

Human near-far

	SIR	SKUR	SPUR	STIR
SIR	5	12	0	3
SKUR	1	12	3	4
SPUR	1	14	5	0
STIR	2	4	3	11

Model near-far

Human far-far

	SIR	SKUR	SPUR	STIR
SIR	11	3	2	4
SKUR	3	12	1	4
SPUR	1	10	7	2
STIR	5	5	1	9

Model far-far

	SIR	SKUR	SPUR	STIR
SIR	18	0	0	2
SKUR	3	15	0	2
SPUR	7	2	10	1
STIR	8	1	1	10

	SIR	SKUR	SPUR	STIR
SIR	16	1	1	2
SKUR	0	16	0	4
SPUR	2	1	14	3
STIR	1	0	0	19

- The confusion matrices show that for human listeners, a far context generally reduces confusions (particularly STIR->SIR).
- The model shows a different pattern of behaviour; SIR->SKUR confusions are reduced but a far context does not substantially improve identification of the consonant.

Conclusions and Future Work

- The effect of reverberation on a consonant identification task has been assessed for human listeners and an ASR system.
- Human listeners use information about the preceding speech context to effect compensation for a reverberated test word; conventional ASR systems do not.
- A computer model in which efferent suppression is mediated by the dynamic range of the preceding context shows limited perceptual compensation.
- Future work will focus on frequency-dependent efferent suppresion in the computer model.
- We will extend this paradigm to study a wider range of consonant confusions.

References

Beeston, A. V. & Brown, G. J. (2010) Perceptual compensation for effects of reverberation in speech identification: A computer model based on auditory efferent processing. Proc Interspeech, Makuhari.

Ferry, R.T. & Meddis, R. (2007) A computer model of medial efferent suppression in the mammalian auditory system. J Acoust Soc Am, 122(6), 3519-3526.

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 Corpus. Linguistic Data Consortium, Philadelphia.

Acknowledgements

Supported by EPSRC grant EP/G009805/1. Thanks to Hynek Hermansky, Tony Watkins and Simon Makin for helpful suggestions and to Ray Meddis for the DRNL program code.