ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Perceptual compensation for the effects of reverberation on consonant identification: A comparison of human and machine performance

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

[1]Department of Computer Science, University of Sheffield, UK
[2]Department of Computer and Information Science, Aalto University, Finland
g.brown@dcs.shef.ac.uk, a.beeston@dcs.shef.ac.uk, kalle.palomaki@aalto.fi

10.21437/Interspeech.2012-469

Abstract

Human listeners are able to perceptually compensate for the effects of reverberation on speech recognition, by exploiting information gleaned from prior exposure to the reverberant environment. We present a computer model of perceptual compensation for reverberation implemented within a hidden Markov model speech recogniser, in which different reverberant speech models are selected depending on the acoustic context preceding a test word. During decoding, observation state likelihoods were computed from two reverberant acoustic models in parallel, and weighted according to the amount of reverberation in the first 500 ms of each utterance. The confusions made by the computer model closely corresponded with those made by listeners in a consonant identification task, and showed a perceptual compensation effect.

Index Terms: reverberation, speech perception, computer model

1. Introduction

Room reverberation presents a challenging problem for automatic speech recognition (ASR) systems, whereas human speech perception is remarkably robust under the same conditions. It has been suggested that mechanisms of 'perceptual compensation' for the effects of reverberation underlie this ability of human listeners. For example, Watkins has shown that listeners use information about the preceding context of a reverberated test word to help them identify it [1]. His experiments focused on one particular speech identification task which employed a synthesized continuum between the words 'sir' and 'stir'. The effect of reverberation on the test word and surrounding context was quantified by measuring the shift in the category boundary between the two words. The /t/ in a reverberated test word 'stir' was more likely to be identified if the preceding context speech was similarly reverberated.

The current paper makes two new contributions. First, we extend Watkins' experiment to natural speech (rather than a synthesized continuum) and a wider range of talkers (20) and consonants (/p/, /t/, /k/). An experiment is described in which perceptual compensation is quantified in terms of the pattern of consonant confusions made by listeners when reverberation is added to the test word and surrounding speech context. Second, we describe a computer model of perceptual compensation that replicates the pattern of confusions observed in the perceptual data. Conceptually, we regard perceptual compensation as a process in which human listeners dynamically weight the evidence from different speech models, depending on the properties of the preceding acoustic context. A concrete implementation of this model is presented, in which observation state likelihoods from hidden Markov models (HMMs) trained under different reverberation conditions are combined prior to Viterbi decoding.

1.1. Perceptual experiment

Forty participants responded to test material (80 utterances) selected from the Articulation Index corpus [2]. Each utterance was of the form CW1 CW2 TEST CW3, where the context words (CW) were drawn from a limited set and the test word (TEST) was 'sir', 'skur', 'spur' or 'stir'. The CW and TEST portions were independently convolved with the left-channel of Watkins' room impulse recordings at 'near' (0.32 m) or 'far' (10 m) source-receiver distance, as described in [1]. This gave the impression of speech at different positions in a room (an L-shaped office with volume 183.6 m^3), with matched or mismatched *context-test* reverberation distance conditions: near-near, near-far and far-far. The early (50 ms) to late ratio was 18 dB at the 'near' distance and 2 dB at the 'far' distance. The utterances were also low-pass filtered (8th order, Butterworth) at five cut-off frequencies between 1 and 4 kHz to investigate frequency-dependent behaviour in the compensation effect.

Stimuli were presented monaurally (to the left ear) via Sennheiser HD480 headphones at a peak level of 48 dB SPL in a sound-isolating booth, matching the conditions used in [1]. Listeners identified the test word with a click of the computer's mouse, which they positioned while looking through the booth's window at the 'sir', 'skur', 'spur' and 'stir' alternatives on the computer

screen. This click initiated the next trial, presented using an iMac, Matlab Version 7.5 (R2007b) and an M-Audio Firewire Audiophile sound interface.

Confusion matrices obtained with the filter cut-off at 4 kHz, the condition at which perceptual compensation for reverberation was most apparent, are presented in the left column of Table 1. Consistent with the previous findings of Watkins [1], increased reverberation on the test word alone caused an increased number of confusions at near-far compared to near-near, notably 'spur' and 'stir' being reported as 'sir'. However, these confusions were largely resolved when the context speech was also reverberated at the far distance (far-far condition). For comparison with Watkins' results, the number of responses for 'sir' and 'not sir' were examined at near-far and far-far distances, resulting in a significant chi-squared value, corrected for multiple tests using the Bonferroni correction, with χ^2=18.729, df=1, and p<0.001.

1.2. Conceptual model

Listeners adapt to reverberant environments, resulting in improved localisation from long-term learning effects (over hours) [3], and decreased speech reception thresholds from short-term echo-suppression effects (over seconds) [4]. Binaural cues are crucial for reverberant listening concerned with localisation or spatial release from masking, however sizeable monaural effects have also been reported for speech perception tasks [1], [4].

Here, we consider a high-level conceptual model of perceptual compensation, in which (fast-acting, monaural) compensation arises from a process of dynamic acoustic model selection. Our perceptual experiment found that speech identification performance is good when the reverberation conditions of the context words and test word are matched (at near-near or far-far). However, a mismatch (at near-far) results in poorer performance since the test word acoustics are incongruous with those of the preceding speech context. This result can be explained by a scheme in which an acoustic model appropriate to the reverberation conditions of the context is selected, and used to decode the test word. We assume that listeners estimate the amount of reverberation present by monitoring the temporal envelope of the context. A mismatch between context and test region reverberation implies that an incorrect acoustic model will be engaged for the test word, and thus more consonant confusions will be made.

2. Computer model

The conceptual model described above was implemented using acoustic model selection in an HMM ASR system. The system was bootstrapped by training it on the TIMIT corpus, which has detailed phonetic transcriptions. The TIMIT phone labels were reduced in number to 39 as described by [5], to give a phone set that could easily be mapped to the labels used in the Articulation Index corpus [2]. HTK [6] was then used to train 39 phone models and a silence model on the TIMIT data. The acoustic models were adapted to the Articulation Index corpus by performing five passes of embedded training on the test signals used in the perceptual experiment.

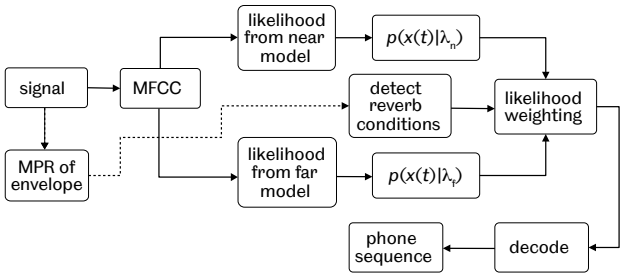

Figure 1: Decoding in the computer model. Observation state likelihoods from 'near' and 'far' acoustic models are weighted according to the reverberation conditions in the preceding acoustic context, detected from the mean-to-peak ratio (MPR) of the speech temporal envelope.

To avoid a mismatch with the experimental stimuli, all training data was low-pass filtered to a cutoff of 4 kHz as described in Section 1.1. Additionally, each training utterance was convolved with 'near' and 'far' room impulse responses to give two reverberated versions. Mel-frequency cepstral coefficients (MFCCs) were then derived for each reverberated utterance. The first 12 MFCCs were used, together with their first- and second-order temporal differences, giving 36 element feature vectors. The feature vectors for the 'near' and 'far' versions of each utterance were concatenated for training in order to provide the likelihood weighting scheme with matching state segmentation for 'near' and 'far' models, and were subsequently split into separate 'near' and 'far' models for decoding.

2.1. Combining likelihood streams

The decoding process is shown schematically in Figure 1. MFCC features were derived from the same experimental stimuli that were presented to listeners, and recognised using the 'near' and 'far' acoustic models in parallel. More specifically, for each feature vector $x(t)$ at time t, the observation state likelihoods $p(x(t)|\lambda_n)$ and $p(x(t)|\lambda_f)$ were computed, where λ_n and λ_f denote the acoustic models trained on 'near' and 'far' reverberated speech respectively. The combined near-far observation state likelihood $p(x(t)|\lambda_{n,f})$ was then obtained as a weighted sum of the 'near' and 'far' likelihoods in the logarithmic domain,

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

where the likelihood weighting factor α is selected according to the reverberation condition (for example, if a 'far' reverberated context is detected, then α is set to zero and subsequent decoding is based on the likelihood from the 'far' acoustic model only).

The combined observation state likelihoods $p(x(t)|\lambda_{n,f})$ were decoded to give a phone sequence using the Viterbi algorithm, as implemented in CTK [7]. Since participants in the perceptual experiment were only required to identify 'sir', 'skur', 'spur' or 'stir', the ASR system was constrained in a similar way; the context words were fixed via semi-forced alignment, and the recogniser identified the test word only.

2.2. Determining the likelihood weighting factor

Two approaches were used to determine the likelihood weighting factor α. In one approach (the oracle condition), it was assumed that the degree of reverberation applied to the context speech was known, and α was set accordingly (i.e., $\alpha = 1$ for a 'near' context and $\alpha = 0$ for a 'far' context). The oracle approach allows an upper bound on the performance of the model to be estimated (i.e., the performance of a compensation mechanism that is provided with perfect information about the context reverberation).

Clearly, for a fully autonomous model the value of α must be estimated from the context speech. Here, this was achieved by classification based on the mean-to-peak ratio (MPR) of the context speech temporal envelope, defined as

$$MPR = \frac{1}{T}\sum_{t=1}^{T} e(t) / \max_{t}[e(t)] \qquad (2)$$

where $e(t)$ is the Hilbert envelope of the first T = 500 ms of the context speech. When more reverberation is present, the mean value of the temporal envelope increases while its peak value remains approximately the same, leading to an increase in the MPR.

The distribution of MPR values for 'near' and 'far' reverberated speech was found to be approximately Gaussian, and therefore a simple Gaussian classifier was adequate to determine the reverberation condition. Given a mean-to-peak ratio value m to be classified, the posterior probability ratio assuming equal priors was computed as

$$d = -\frac{1}{2}\left[\begin{array}{r}(m-\mu_n)^2/\sigma_n^2 - (m-\mu_f)^2/\sigma_f^2 \\ + \ln\sigma_n^2 - \ln\sigma_f^2\end{array}\right] \qquad (3)$$

where μ_n and μ_f represent the mean MPR of 'near' and 'far' reverberated test signals respectively, and σ_n and σ_f are the corresponding standard deviations. If $d \geq 0$ then the context speech was classified as 'near' and we set $\alpha = 1$, otherwise $\alpha = 0$. The classifier performance was 83% correct across all reverberation conditions in the test set.

3. Results

To quantify the similarity between the consonant confusions made by human listeners and the computer model, we follow the approach described by [8]. Each row of the human and model confusion matrices are compared in a 2×4 contingency table using Pearson's phi-squared (ϕ^2) index. A value of $\phi^2 = 1$ indicates complete dissimilarity of a set of frequencies (i.e., the frequency distributions are non-overlapping) whereas a value of $\phi^2 = 0$ indicates equality.

Results obtained using oracle likelihood stream selection are shown in the middle column of Table 1. There is a good match to human performance ($\phi^2 < 0.1$ in all conditions). Qualitatively, the main effects visible in the human data are reproduced by the model; few confusions occur in the near-near condition, but there are frequent confusions in the near-far condition. For both human listeners and the computer model, the most common confusion is 'stir' reported as 'sir'. In the far-far condition, these confusions are largely resolved both for listeners and the computer model, indicating a 'perceptual compensation' effect.

The right column of Table 1 shows results obtained using the fully autonomous version of the computer model, in which the likelihood weighting factor α is obtained from a classifier operating on the MPR of the context speech envelope. Again, few confusions occur in the near-near condition, and the confusions made by the model in the near-far condition closely resemble those made by human listeners ($\phi^2 < 0.1$ in all conditions).

4. Discussion and Conclusions

A listening experiment has been presented that provides further evidence of perceptual compensation for the effects of reverberation in human speech perception. The experiment extends Watkins' findings for a single stop consonant [1] by showing that compensation is apparent in the perception of three stop consonants. Moreover, perceptual compensation has been demonstrated for a relatively natural speech identification task, in which the utterances varied in length and context words, and were produced by both male and female talkers.

A computer model of the perceptual compensation effect based on acoustic model selection gave a close match to the pattern of consonant confusions made by human listeners. The premise of the model is simple; acoustic cues in the speech preceding the test word determine the acoustic model that is engaged when the test word is decoded. If there is a mismatch between the reverberation conditions applied to the test word and context words, consonant confusions increase. A version of the model that determines the reverberation condition from the MPR of the context speech envelope gave similar confusion patterns to a version of the model in which the re-

Human near-near					Oracle model near-near						MPR model near-near					
	SIR	SKUR	SPUR	STIR		SIR	SKUR	SPUR	STIR	ϕ^2		SIR	SKUR	SPUR	STIR	ϕ^2
SIR	37	0	0	3	SIR	32	0	0	8	0.0329	SIR	32	0	0	8	0.0329
SKUR	0	40	0	0	SKUR	0	38	0	2	0.0256	SKUR	0	38	0	2	0.0256
SPUR	0	1	38	1	SPUR	2	0	38	0	0.0500	SPUR	2	0	34	4	0.0628
STIR	0	0	0	40	STIR	0	2	0	38	0.0256	STIR	2	2	2	34	0.0811
Human near-far					Oracle model near-far						MPR model near-far					
	SIR	SKUR	SPUR	STIR		SIR	SKUR	SPUR	STIR	ϕ^2		SIR	SKUR	SPUR	STIR	ϕ^2
SIR	37	0	0	3	SIR	36	2	2	0	0.0877	SIR	36	0	2	2	0.0277
SKUR	6	29	2	3	SKUR	6	34	0	0	0.0675	SKUR	6	34	0	0	0.0675
SPUR	16	3	19	2	SPUR	6	2	30	2	0.0902	SPUR	10	2	28	0	0.0664
STIR	16	2	1	21	STIR	18	6	0	16	0.0474	STIR	16	6	0	18	0.0404
Human far-far					Oracle model far-far						MPR model far-far					
	SIR	SKUR	SPUR	STIR		SIR	SKUR	SPUR	STIR	ϕ^2		SIR	SKUR	SPUR	STIR	ϕ^2
SIR	33	1	1	5	SIR	22	4	4	10	0.0933	SIR	28	2	4	6	0.0329
SKUR	0	34	0	6	SKUR	2	36	0	2	0.0507	SKUR	4	32	0	4	0.0558
SPUR	3	2	31	4	SPUR	4	0	36	0	0.0814	SPUR	6	0	32	2	0.0460
STIR	2	1	0	37	STIR	0	0	0	40	0.0390	STIR	0	0	0	40	0.0390

Table 1: Confusion matrices for human listeners (left), computer model given oracle information about the reverberation condition (center) and computer model that uses a MPR metric to determine the reverberation condition (right). Reverberation conditions are labelled as *context-test* distance. Rows correspond to the stimuli presented; columns record the responses.

verberation condition was known *a priori*.

The model could be extended in a number of ways. Firstly, the MPR is a crude measure of the amount of reverberation present, and more sophisticated metrics could be used that are also perceptually plausible (e.g., identification of reverberation tails, or reverberation measures based on the smearing of harmonic structure). Secondly, since recent experiments by Watkins [9] suggest that perceptual compensation is underlain by a mechanism that works within individual frequency bands, we will develop a channel-independent model in the future.

Finally, we note that the model of perceptual compensation described here is closely related to some approaches to noise-robust and reverberation-robust ASR. For example, [10] proposes an ASR system in which blind estimation of reverberation time is used to select one of several acoustic models, in a similar way to the scheme proposed here. Our model therefore represents a convergence of ideas from speech technology and psychoacoustics, and is a first step towards our eventual goal: to develop computer models that closely replicate human performance in speech perception, but also serve as practical ASR systems.

5. Acknowledgements

GJB and AVB were supported by EPSRC research grant G009805/1. KJP was supported by the Academy of Finland (136209) and the IST Programme of the European Community, under the PASCAL2 Network of Excellence (IST-2007-216886).

6. References

[1] A. Watkins, "Perceptual compensation for effects of reverberation in speech identification," *J. Acoust. Soc. Am.*, vol. 118, no. 1, pp. 249–262, 2005.

[2] J. Wright, "Articulation index," Linguistic Data Consortium, Philadelphia, Tech. Rep., 2005.

[3] B. Shinn-Cunningham, "Learning reverberation: Considerations for spatial auditory displays," in *Proc. 2000 Int. Conf. on Auditory Display*, Atlanta, GA, 2000.

[4] E. Brandewie and P. Zahorik, "Prior listening in rooms improves speech intelligibility," *J. Acoust. Soc. Am.*, vol. 128, no. 1, pp. 291–299, 2005.

[5] K. Lee and H. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 1641–1648, Nov. 1989.

[6] *HTK version 3.4*. http://htk.eng.cam.ac.uk, 2011.

[7] *CTK version 1.3.5*. http://tinyurl.com/6yard2w, 2011.

[8] T. Jürgens and T. Brand, "Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model," *J. Acoust. Soc. Am.*, vol. 126, no. 5, pp. 2635–2648, 2009.

[9] A. J. Watkins, S. J. Makin, and A. P. Raimond, "Constancy in the perception of speech when the level of room reflections varies," in *Binaural processing and spatial hearing*, J. Buchholz, T. Dau, J. Dalsgaard, and T. Poulsen, Eds. Ballerup, Denmark: Danavox Jubilee Foundation, 2010, pp. 371–380.

[10] L. Couvreur and C. Couvreur, "Blind model selection for automatic speech recognition in reverberant environments," *Journal of VLSI Signal Processing*, vol. 36, pp. 189–203, 2004.