A glimpsing model of speech perception

Martin Cooke
Department of Computer Science, University of Sheffield, UK
m.cooke@dcs.shef.ac.uk

ABSTRACT

Do listeners utilise glimpses of a clean speech target in noisy backgrounds? This paper presents a model of speech perception motivated by the glimpses notion. Using a corpus of VCV sequences, listeners' identification performance is measured in a number of noise conditions chosen to vary both in global SNR and in the number and distribution of glimpses they afford. Subjects' performance is compared to that of a computational model which employs missing data techniques to handle glimpses. A close match to subjects' performance can be obtained by assuming that listeners require available glimpses to occupy a certain minimum extent in time and frequency.

1. INTRODUCTION

We have little idea of the processes which underpin human speech communication in everyday conditions. Most models of speech perception focus on clean speech material, and consequently shed little light on the sorts of representations and processes involved in more realistic conditions. This focus has also led in automatic speech recognition to the widespread adoption of speech parameters such as the truncated cepstrum, which encodes *global* spectral shape, even though it is highly unlikely that complete spectra are accessible in noisy environments.

In a recent review, Assmann & Summerfield [1] enumerate some of the distorting effects present in typical communicative settings – reverberation, additive and channel noise – and go on to describe schemes which the auditory system might use to ameliorate their effects. One such scheme they propose to deal with additive noise is *glimpsing*. The essential idea is that listeners may be able to exploit speech glimpses – spectro-temporal regions where the speech signal is more energetic than the non-speech background sources – in order to identify speech in noise. Glimpsing represents an alternative to strategies which rely on recovering the clean speech signal from a noisy mixture.

A significant body of experiments employing distorted speech lends support to this idea. For instance, manipulations which result in spectral holes [2,3] or temporal gaps [4,5] are routinely handled by listeners. For example, Drullmann [6] demonstrated sentence intelligibility scores of 60% for signals in which 98% of information was missing. The case for glimpsing is argued at greater length in [7]. Here, we highlight two simulations which provide additional support for glimpsing.

Figure 1 plots the nonstationarity of a noise corpus as a function of the gain in SNR resulting from the application of a high-performance noise estimation algorithm [8]. Nonstationarity is defined here as the mean (across frequency) of energy variance in time, measured in an auditory spectrographic representation. Figure 1 demonstrates that, for the current generation of noise estimators at least, nonstationarity is harmful, and that a highly nonstationary 'noise' consisting of a single interfering talker cannot be handled effectively.

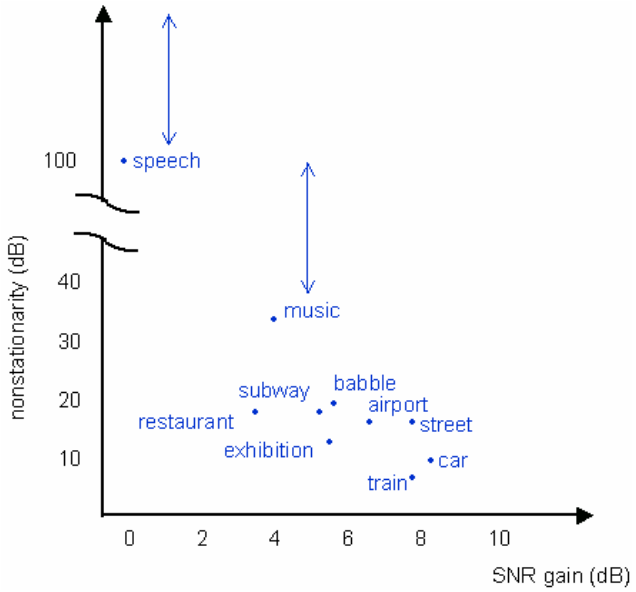

Figure 1: Gain in signal-to-noise ratio obtained by a high-performance noise estimation algorithm versus a measure of nonstationarity for a variety of noises.

So, how does the curse of nonstationarity affect listeners' performance? Miller [9] answered this question by measuring the accuracy with which listeners identified monosyllables as a function of SNR for noise backgrounds consisting of a mixture of N speakers. For increasingly larger N, the mixture tends towards the (stationary) average long-term speech spectrum. Miller found an advantage of up to 12 dB (measured in terms of Speech Reception Threshold) for the single speaker background as opposed to a background of 8 speakers, demonstrating that, for listeners, nonstationarity of the noise background is highly beneficial. Similar findings were reported by Festen & Plomp [10] using speech and steady-state noise maskers.

ISBN 1-876346-48-5

Could it be that listeners are using the increased opportunity for glimpses afforded by nonstationarity? Figure 2 provides an indication of the potential of a glimpsing explanation for the case of an n-speaker background, for n=1 and 8. Here, using an auditory spectrographic representation, the percentage of frequency regions which contain more energy than the noise background is plotted against the global SNR of the mixture. Results represent an average across a large corpus of speech. It is clear from this figure that, at any SNR, substantially more glimpses of the target speech are available when the background is nonstationary (1 speaker) than when it is nearly stationary (8 speakers).

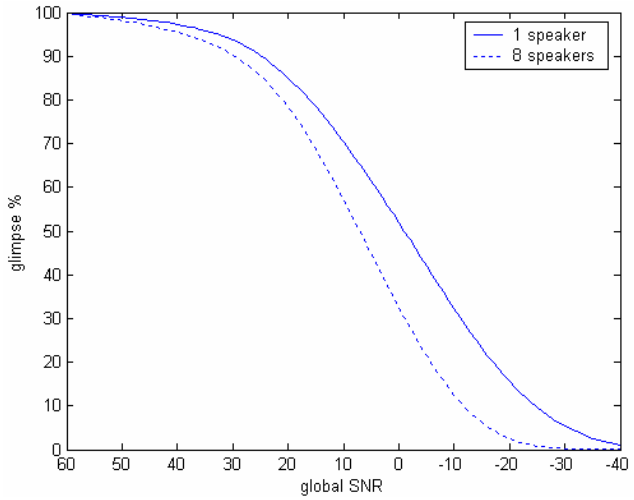

Figure 2: Proportion of glimpses of a speech target available in the presence of a background composed of 1 or 8 speakers, as a function of SNR.

Two studies have compared listeners' performance with that of a glimpsing model. Lippmann & Carlson [11] showed qualitative similarities between listeners and model in identifying low and high-pass filtered material. De Cheveigné & Kawahara [12] investigated a missing-data model for the perception of synthetic vowels. Both studies demonstrated the potential of a glimpsing model, but were unable to provide detailed insights into its operation. Speech material used in the model in [11] differed from that of the listeners with which it was compared, while the stationary vowels used in [12] are not representative of speech in general. Second, neither study investigated different definitions of what might constitute a glimpse.

The purpose of the current study was to provide a direct comparison between listeners and a glimpsing model using a common corpus, and to determine how differing glimpse definitions affect model performance. For example, it is unclear how dominant a time-frequency region has to be to qualify as a glimpse. Further, it is possible that glimpses cannot be detected unless they occupy a certain minimum extent in time and frequency.

Section 2 describes the stimuli employed in this investigation, while sections 3 and 4 report on listeners' and model performance in an identification experiment.

2. SPEECH AND NOISE MATERIALS

2.1 Speech

A key requirement for test materials was their ability to support both perception testing and model training. A further requirement was that materials should be reasonably representative of natural speech, yet avoid the engagement of higher-level cognitive factors (lexical and above) in their processing. A VCV corpus collected by Shannon et al [13] meets these requirements. This corpus contains VCV and CV tokens for 3 vowels and 25 consonants, spoken by 5 male and 5 female talkers. Ten repetitions of each token were made. Consequently, although the corpus was collected for perceptual testing, it can also be used for training whole-word HMMs.

The current study used a subset of the Shannon corpus consisting of VCVs from the 5 male speakers with V = /a/ and the 16 Cs /b, d, g, p, t, k, m, n, l, r, f, v, s, z, ʃ,tʃ/. The subset was chosen to provide a reasonably difficult yet manageable task for subjects lacking an extensive background in phonetics. Consonants such as /θ/ and /ð/ with a confusable orthographic representation were omitted.

Of the 10 repetitions of each token by each of the 5 male speakers, 8 were used for model training (giving 40 exemplars per token). The remaining 2 repetitions of each token made up the test set, a total of 160 items.

2.2 Background 'noise'

The two noise conditions shown in figure 2 were employed in this study, namely multispeaker babble for n=1 and 8. These were chosen since previous listening studies [9,10] have shown that they lead to distinct differences in identification performance. Multispeaker babble was created by recording long segments of speech, equalising levels and adding n individual speech signals. The resulting babble signals were then reversed to reduce the risk of linguistic interference. Randomly-selected babble segments were added to speech tokens at 3 SNRs (0, -6, -12 dB). In a pilot experiment, these SNRs were demonstrated to provide a useful performance range.

3. LISTENERS

3.1 Procedure

Two groups of 6 subjects took part in the experiment, which consisted of 2 sessions several days apart. One group underwent the 8-speaker babble condition in the 1st session, followed by the 1-speaker case. The order of conditions was reversed for the 2nd group to eliminate order effects. Prior to the 1st session, each subject was screened by testing with 5 noise-free repetitions of each of the 16 /aCa/ tokens. A target performance of 95% was set for the pretest. One subject failed to meet this target and was substituted.

Within each session, the experiment was split into 3 blocks by noise level (0, -6 and –12 dB). Within each subject group, individuals received the 3 blocks in a different sequence, again to remove ordering effects. Each block consisted of 192 tokens. The initial 32 tokens were treated as practice stimuli and ignored, although subjects were not

ISBN 1-876346-48-5

told this. The remaining 160 tokens constituted the test set as described in section 2. These tokens were presented in a randomised order. Each block lasted 5-10 minutes.

Stimuli were presented at 24 kHz under computer control over Sennheiser HD250 headphones in an IAC single-walled soundproof room at 70dB SPL. Subjects used a cordless mouse to supply responses, selecting one of 16 buttons on a screen located outside the booth.

3.2 Results

Figure 3 summarises listeners' identification performance as a function of SNR and noise condition. These results are comparable with those obtained by Miller [9], showing an 8-12 dB advantage for the single-speaker noise background over the 8-speaker case.

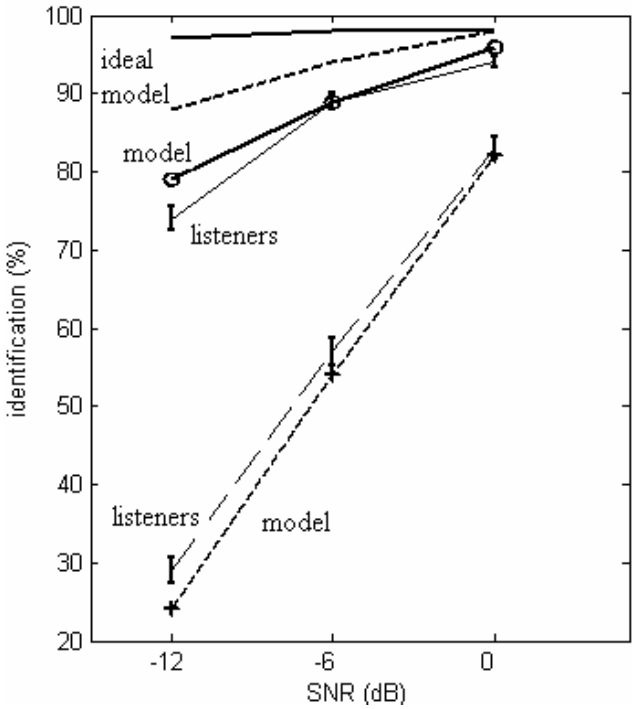

Figure 3: Listeners' and model performance as a function of SNR, in single-speaker (solid line) and 8-speaker (dotted line) babble. Ideal model performance is also shown.

4. MODEL

4.1 Procedure

A set of 16 isolated-word HMMs was used to model the /aCa/ corpus. Each HMM had 8 states, with a density model consisting of a 4-component Gaussian mixture in each state. Speech parameters were successive 10 ms frames of smoothed energy at the output of a gammatone filterbank [14] consisting of 40 filters equally spaced on an ERB-rate scale from 50-7500 Hz. The test set was identical to that heard by listeners. Performance in clean speech was 99%.

For the noisy conditions, information in data vectors was partially-specified to simulate glimpsing. Incomplete data vectors were recognised using bounded marginalisation, a missing data technique described in detail in [15].

Two sets of simulations were carried out. In the first, the effect of varying the threshold for glimpse detection was evaluated. The second simulation explored the effect of specifying a minimum spectro-temporal extent.

4.2 Variation in glimpse detection threshold

The glimpse detection threshold can be defined as the amount (in dBs) by which the target speech energy must exceed that of the noise background, for each time-frequency region. For illustration, figure 4 (top) shows those glimpses which result from a 0 dB detection threshold. It is clear from this figure that glimpses are not randomly scattered in time-frequency but tend to occupy regions surrounding speech formants, and that most of the glimpses correspond to vowel portions of the VCV token.

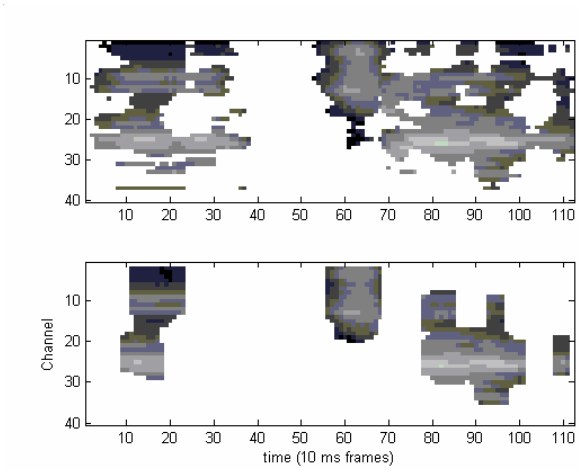

Figure 4: Glimpsing in an auditory spectrogram. Top: Regions with a locally positive SNR. Bottom: Regions which additionally possess more than a minimum extent in time (40 ms) and frequency (9 channels).

The optimum performance of the model occurs for glimpse detection thresholds in the range –2 to 2 dB. Figure 3 ('ideal model') shows consonant identification accuracy for a threshold value of 0 dB. Clearly, the model significantly outperforms listeners in every condition. No single glimpse detection threshold in the range tested (–6 to 14 dB) provided a close fit to the pattern of listeners' performance. These results suggest that there is sufficient information in glimpses to support speech perception in noise, but that listeners do not have access to the rather ideal information used in the simulations. Figure 4 (top) shows that some glimpses are rather short and/or occupy a very narrow frequency region. How realistic is it to assume that listeners can detect, and label as 'speech' such brief glimpses in a sea of noise? The second set of simulations tested the possibility that listeners require glimpses to possess a minimum extent in time and frequency.

4.3 Variation in minimum glimpse extent

Glimpse extent can be varied by pruning those spectro-temporal fragments which do not belong to a rectangular region of a certain minimum extent. Figure 4 (bottom) illustrates this process for a region of 6.3 ERBs (9 channels)

ISBN 1-876346-48-5

and 40 ms (4 frames). A search was conducted over glimpse extents from 1-100 ms and 1-20 channels for a range of detection thresholds. The best fit to listeners' performance was found for a minimum glimpse extent of 9 channels and 40 ms, at a detection threshold of 0 dB. Figure 3 shows model performance in this condition.

4. DISCUSSION

A model which requires a clear view of the speech target over a frequency range of 5.5-7.0 ERB and a duration of 30-50 ms can explain the overall identification performance of listeners in this task. The frequency extent represents a division of the range of importance for speech into 4 or 5 equally-spaced bands. Such a model is compatible with measurements of listeners' ability to use information in different frequency regions [16].

However, the close match should be interpreted with caution, since it is the result of an optimisation over a number of free parameters. It is possible that a model employing a different speech representation or glimpse definition would produce similar results. Further insight into the model can be obtained by examining the pattern of confusions produced by listeners and the model. Such an analysis shows that there are many tokens for which a majority of subjects agree on an (erroneous) response. However, the model shows little agreement with subjects at this level of detail.

The auditorily-motivated representation used in this study takes into account simultaneous masking, but nonsimultaneous masking is not present. Consequently, it is likely that glimpses provided by the simulation do not accurately represent the information available to listeners. Similarly, the notion that a useful glimpse consists simply of a sufficiently-large, unobstructed view of the target signal is oversimplified. It is possible that listeners use auditory organisation principles such as grouping by onset synchrony to determine which components are part of the same source [17].

5. CONCLUSIONS

At any given SNR, listeners perform better at identifying consonants in vowel contexts when those sounds are presented in nonstationary backgrounds than in stationary backgrounds. This is in striking contrast to the deleterious effect that nonstationarity has on current noise estimation schemes. A model based on glimpsing speech is compatible with these findings, and with suitable restrictions on what constitutes a glimpse, can explain listeners' overall identification performance. However, more work is necessary to improve the glimpsing criterion and to compare performance with that of non-glimpsing models on the same data.

ACKNOWLEDGEMENTS

Thanks to Professor R Shannon for the VCV corpus used here and to Sarah Simpson for the SNR gains in Figure 1.

REFERENCES

[1] P. Assmann and Q. Summerfield, "The perception of speech under adverse conditions", to appear in *Speech Processing in the Auditory System* (eds Greenberg, Ainsworth, Popper & Fay) Springer.

[2] R.P. Lippmann, "Accurate consonant perception without mid-frequency speech energy", *IEEE Trans. Speech & Audio Proc.*, **4**, pp. 66-69, 1996.

[3] K. Kasturi, P.C. Loizou, M. Dorman & T. Spahr, "The intelligibility of speech with "holes" in the spectrum", *J. Acoust. Soc. Am*., **112**, pp. 1102-1111, 2002.

[4] G.A. Miller and J.C.R. Licklider, "The intelligibility of interrupted speech", *J. Acoust. Soc. Am.*, **22**, pp. 167-173, 1950.

[5] W. Strange, J.J. Jenkins and T.L. Johnson, "Dynamic specification of coarticulated vowels", *J. Acoust. Soc. Am.*, **74**, pp. 695-705, 1983.

[6] R. Drullmann, "Speech intelligibility in noise*", J. Acoust. Soc. Am*., **98**, pp. 1796-1798, 1995.

[7] M.P. Cooke, "Glimpsing speech", *J. Phonetics*, to appear in 2003.

[8] J. Sohn, N.S. Kim and W. Sung, "A statistical model-based voice activity detection", *IEEE Signal Processing Letters*, **6**, pp. 1-3, 1999.

[9] G.A. Miller, "The masking of speech", *Psych. Bull*., **44**, pp. 105-129. 1947.

[10] J.M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing", *J. Acoust. Soc. Am*., **88**, pp. 1725-1736, 1990.

[11] R. P. Lippmann and B. Carlson, "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise", *Proc. Eurospeech*, 1997.

[12] A. de Cheveigné and H. Kawahara, "Missing-data model of vowel identification", *J. Acoust. Soc. Am*., **105**, pp. 3497-3508, 1999.

[13] R.V. Shannon, A. Jensvold, M. Padilla, M.E. Robert and X. Wang, "Consonant recordings for speech testing", *J. Acoust. Soc. Am*., **106**, pp. L71-L74, 1999.

[14] M.P. Cooke, *Modelling Auditory Processing and Organisation*, Cambridge University Press, 1993.

[15] M.P. Cooke, P.D. Green, L. Josifovski and A. Vizinho, "Robust automatic speech recognition with missing and uncertain acoustic data", *Speech Communication*, **34**, 267-285, 2001.

[16] P.A. Howard-Jones and S Rosen, "Uncomodulated glimpsing in 'checkerboard' noise", *J. Acoust. Soc. Am*., **93**, pp. 2915-2922, 1993.

[17] M.P. Cooke and D.P.W. Ellis, "The auditory organisation of speech and other sources in listeners and computational models", *Speech Communication*, **35**, pp. 141-177, 2001.