Model-driven detection of clean speech patches in noise

Jonathan Laidler[1], *Martin Cooke*[1], *Neil D. Lawrence*[2]

[1]Department of Computer Science, University of Sheffield, Sheffield, UK
[2]School of Computer Science, University of Manchester, Manchester, UK
{j.laidler, m.cooke}@dcs.shef.ac.uk, neill@cs.man.ac.uk

Abstract

Listeners may be able to recognise speech in adverse conditions by "glimpsing" time-frequency regions where the target speech is dominant. Previous computational attempts to identify such regions have been source-driven, using primitive cues. This paper describes a model-driven approach in which the likelihood of spectro-temporal patches of a noisy mixture representing speech is given by a generative model. The focus is on patch size and patch modelling. Small patches lead to a lack of discrimination, while large patches are more likely to contain contributions from other sources. A "cleanness" measure reveals that a good patch size is one which extends over a quarter of the speech frequency range and lasts for 40 ms. Gaussian mixture models are used to represent patches. A compact representation based on a 2D discrete cosine transform leads to reasonable speech/background discrimination.

Index Terms: speech separation, glimpsing, model-driven, spectro-temporal patches.

1. Introduction

Listeners are able to identify speech across a wide range of adverse conditions [1, 2], even in cases where the global signal-to-noise ratio (SNR) is very low. Listeners appear to perform some kind of source separation to isolate components of a target voice, prior to, or in conjunction with, recognition of the target speech.

Previous attempts to model source separation have used either source-driven (bottom-up) or model-driven (top-down) processes. Auditory scene analysis [3] has inspired computational approaches [4, 5] which can be viewed as source-driven techniques since they look for evidence in primitive auditory features for source properties such as fundamental frequency or spatial location. Other source-driven approaches work by exploiting statistical independence of the acoustic sources in a mixture [6], but this is problematic when more sources than sensors exist. Purely model-driven approaches [7, 8] attempt to find the combination of learnt models which best describe the observations. These rely on the existence of models for each noise source, and can be computationally expensive when dealing with the arbitrary combinations of models needed to decode real acoustic environments.

Recent studies of speech in noise have shown that "glimpses", or spectro-temporal regions where the target speech is dominant, contain more than enough information to serve as a basis for speech perception [9], even when such regions make up only a fraction of a noisy acoustic scene. This is due largely to the sparse and redundant nature of speech when viewed as a time-frequency signal. By exploiting the glimpsing phenomenon, models of source separation can be simplified since the system is no longer required to fully segregate the target speech from the background. Rather, the task is one of locating an unlabelled set of glimpses that are dominated by one source or another. Regions where no source dominates are simply ignored, and the unlabelled glimpses can be passed to a speech fragment decoder [10] which finds the optimal subset of glimpses at the same time as determining the most likely speech interpretation. Speech segregation is then a side-effect of recognition, and no detailed models of the (one or more) sources which constitute the background are required.

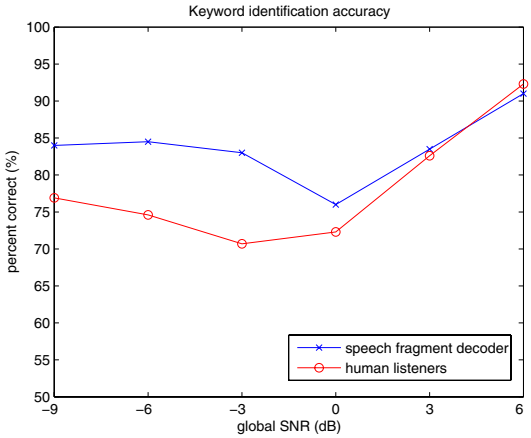

Figure 1: *Comparison of listeners with a speech fragment decoder provided with* a priori *fragments, evaluated on a keyword identification task.*

The viability of the glimpsing approach can be illustrated by measuring the upper limit on the performance of a speech fragment decoder when supplied with a set of glimpses known to be "correct" in the sense of having been produced with prior knowledge of the local signal-to-noise ratio (SNR) as a function of time and frequency. Pairs of similar sentences from the Grid Corpus (see section 2.1) were added at a range of global SNRs, and fragments were formed based on connected spectro-temporal regions where one or other sentence dominated. These *a priori* fragments were then decoded by a speech fragment decoder [10]. Keyword recognition accuracy is shown in figure 1. Even at the most adverse SNR the score is high (83%) in spite of the fact that in terms of spectro-temporal area only 29% of the target speech has been used. Figure 1 also plots listener performance on the same task [11]. For positive SNRs, the speech fragment decoder performance is essentially identical to that of listeners, suggesting that listeners may well be capable not only of exploiting the limited information in glimpses, but also of identifying an ideal set of spectro-temporal regions dominated by a single source. At adverse SNRs, the decoder outperforms listeners, possibly because the model is not affected by attentional limitations known to degrade human performance when faced with similar target and masking stimuli. An additional factor limiting listener performance may be an inability to identify the maximal set of glimpses.

Our goal is to determine whether spectro-temporal "patches" belonging to a single speaker can be identified using model-driven techniques. The current study tackles two issues. The first question concerns the size of patches best suited to the task of identifying clean speech. Too small a patch size limits the ability of a model to discriminate clean speech from non-

clean speech or non-speech background, while larger patches increase the probability that the region contains information from more than one source. The other issue concerns modelling the distribution of clean speech patches. Two density estimation approaches based on Gaussian mixture models are tested, one using raw spectro-temporal energy, the other employing a more compact representation.

2. Corpus and preprocessing

2.1. Corpus

Additive mixtures of utterances from pairs of talkers were used throughout this study. Sentences were drawn from the Grid Corpus [12] which was designed to support joint computational-behavioural studies of audio and audiovisual speech recognition. The Grid Corpus consists of 1000 utterances spoken by each of 34 talkers (male and female). All utterances have a simple 6-word form as in the phrase "place blue at f nine now". Four colours, 25 spoken alphabet letters and 10 spoken digits act as keywords. The colour keyword is typically used to identify which of a pair of simultaneously-presented utterances should be regarded as the target, and the task for listeners/algorithms is to report the letter-digit combination spoken by the target talker.

Here, pairs of Grid utterances were added at 6 SNRs (6, 3, 0, -3, -6, -9) dB. At each SNR, 100 pairs were constructed with the same target talker (talker 5) while the masking talker varied. In approximately one-third of pairings, the masker was of a different gender; in another third, the gender was the same as the target but the talker was different; in the remaining third, the target and masker sentences came from the same talker.

2.2. Preprocessing

Utterances were parameterised using an auditory "ratemap" representation created as follows. First, the signal was filtered by a bank of 64 gammatone filters whose centre frequencies ranged from 50 to 8000 Hz with equal spacing on an ERB-rate scale. Next, the instantaneous envelope at the output of each filter was extracted using the Hilbert transform and smoothed using a leaky integrator with a time constant of 8 ms, reflecting psychophysical estimates of the auditory temporal window [13]. Finally, the smoothed envelope was integrated into 10 ms frames and log-compressed. Further details of these processes can be found in [4]. An example of the resulting ratemap representation for a mixture of two talkers is given in figure 3.

3. Patch size determination

3.1. Cleanness measure

There is a tradeoff between large patches, which are more likely to contain evidence of more than one source, and small patches, which contain insufficient information to allow speech/non-speech discrimination. In the extreme case of a single "pixel" energy value in the ratemap, the discriminative power is very low indeed.

Here, we make the simplifying assumption that patches are rectangular, anticipating a later patch grouping process which forms larger, non-rectangular regions. It is likely that for large patch sizes the rectangularity constraint is sub-optimal. For example, while it is sometimes the case that parallel vertical edges are produced by synchronous across-frequency energy increases, the detailed edge shape depends on other factors, such as formant transitions in the target and nonstationarity of the masker. Similarly, F0 and formant frequency dynamics across a patch will lead to non-horizontal structures.

To quantify the degree to which a given patch size is likely to represent material from a single source in a mixture, a measure of patch "cleanness" is introduced. Given a mixture m of K sources s^k, the cleanness of a specific rectangular patch with frequency range F (channels) and duration D (frames) is defined as

$$C = \max_{k \epsilon K} \frac{\sum_{t,f} \left(E(m_{tf}) - E(s^k_{tf}) < \epsilon \right)}{F \times D}, \qquad (1)$$

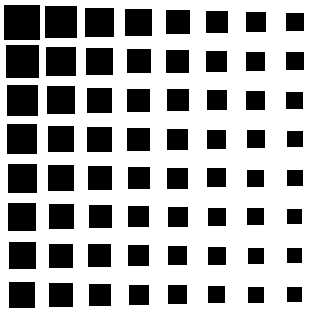

Figure 2: *Hinton diagram where the area of each block represents the percentage of patches considered clean for different sizes in frequency (y-axis) and time (x-axis), increasing in multiples of two. The top left corner represents a 2 x 2 patch while the bottom right corner represents a 16 x 16 patch.*

where $E(m_{tf})$ and $E(s^k_{tf})$ represent the log energy in the ratemaps of the mixture and the individual sound sources respectively at time t and frequency f. The threshold ϵ was chosen to maximise the percentage of spectro-temporal pixels allocated uniquely to a single source, which occurred for $\epsilon = 3.8$dB. At this threshold, 91% of all pixels in a development corpus of two-talker speech mixtures were unambiguously classified as either target or masker.

3.2. Patch size variation and cleanness

Patch sizes ranging from 2 to 16 frequency channels by 2 to 16 time frames were considered. Patches with $C > 0.95$ were regarded as clean (i.e. a single source dominated over 95% of its area). Every patch centred on every pixel in the 100 ratemaps in the 0 dB condition was used in the estimation of cleanness. The proportion of clean patches as a function of patch size is shown in figure 2. While the decrease in cleanness proportion is quite gradual as frequency range increases, the effect of increasing duration is larger.

There is no reason to believe that when choosing a suitable patch size it should be the same for all frequency locations. When each frequency channel was treated independently, it was found that at high frequencies, brief (40 ms) but spectrally-wide (16 channels) patches could be used, while at low frequencies, longer (80 ms) but spectrally-narrow (2 channels) patches showed equivalent cleanness proportions. In the middle range of frequencies, intermediate patch sizes were evident. This is illustrated in figure 3 which shows a ratemap of two sentences mixed at 0 dB global SNR.

4. Patch modelling

4.1. Gaussian mixture models

One approach to discriminating between patches representing a single speech source and those coloured by contributions from other sources is to construct a density model for clean speech patches. Such a model will represent speech in general rather than specific sounds and must allow for the fact that a clean speech patch could conform to one of a large number of patterns. Here, a mixture of Gaussians is used to handle multimodality, and each component models a specific patch pattern.

4.2. Raw energy representation

The simplest form of patch encoding is to use the raw energy values in the patch directly. Such features are not independent, since the energy in adjacent time frames and frequency channels can be strongly correlated. For this reason, full covariance mixture models were used.

A separate Gaussian mixture model (GMM) was trained for patches centred on each frequency channel, because certain patterns may be specific to one frequency location. Models were

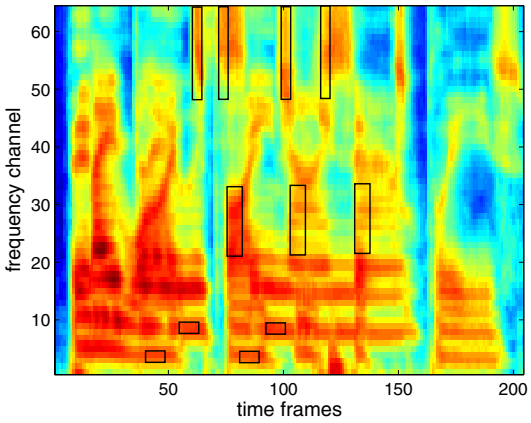

Figure 3: *Patches of clean speech in a two-talker mixture appear in approximately equal measures when the patches are tall and thin at high frequencies and short and wide at low frequencies.*

trained only on sentences spoken by the target talker (talker 5) in order to determine not only how well the system could identify speech from patches, but the extent to which it could identify speech from a specific talker. Five hundred training sentences spoken by the target talker were chosen, with no overlap with the evaluation set of mixtures. Maximum likelihood estimation of model parameters was used, initialised via k-means, followed by expectation maximisation (EM) [14], iterated until convergence. EM does not guarantee a global maximum, so the model with the maximum likelihood on the training set was chosen from 5 random initialisations.

The number of components in the mixture model should relate loosely to the number of different "patterns" observed in patches centred on one frequency channel. The optimum number of components may be frequency dependent. A model selection technique was used to find the number of components which maximised the likelihood of a validation set. Here we used 5-fold cross-validation to train models on patches of size 6 by 4 (24 features) with between 50 and 300 components in multiples of 10. The number of components which provided the highest validation log likelihood for patches centred on each frequency channel is shown in figure 4. The optimum number of components is indeed frequency-specific. While over 200 components are required at low frequencies (up to around 1 kHz), 100 are sufficient at frequencies above 1500 Hz, with a rather narrow transitional zone in between. The difference between high and low frequencies is striking, and seems likely to reflect the interaction of speech harmonics with auditory frequency resolution. As figure 3 illustrates, resolved harmonics appear in the ratemap at low frequencies, while higher frequencies are dominated by unresolved harmonics and instead show formant information.

4.3. Patch likelihoods

Based on the patch size findings in section 3, separate GMMs were trained for patches with sizes up to 12 frequency channels by 8 time frames (larger patches were judged to be too likely to be dominated by more than one source). A uniform patch size was used in each channel. For channels 1-22 GMMs with 200 components were employed while GMMs with 100 components were used for the remaining channels.

Figure 5(a) shows the distribution of log likelihoods for patches of size 12 x 8 identified *a priori* as being dominated by the target speech, the background speech or neither source. The mean log likelihood for clean speech is higher than for either the background speech or the ambiguous regions where both sources contributed, although the distributions overlap.

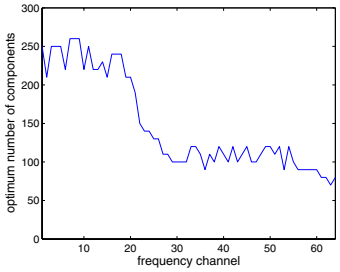

Figure 4: *The number of mixture components yielding the highest likelihood for patches centred on each frequency channel.*

Surprisingly, the mean likelihood of the background speech is no higher than for the ambiguous regions, suggesting that the GMM represents a talker-dependent model for the target. Similar results were found for other patch sizes.

Figure 5(b) shows a breakdown of the mean likelihoods for each frequency channel. At low frequencies, there is a greater difference between the clean target speech pixels and the ambiguous pixels than at high frequencies. It is not clear whether this is caused by the use of a uniform patch size which may have less discriminative power at higher frequencies.

For any successful model, patches with a large positive or negative local SNR should have a high likelihood. Figure 5(c) depicts the relationship between the average local SNR of a patch and its likelihood. The plot was formed by first computing a 2D histogram of average local SNR within a patch versus log likelihood, then normalising so that each local SNR bin (column) had an equal number of entries. This normalisation is necessary since the distribution of local SNRs across patches is not uniform. There is a clear positive correlation when the SNR is greater than zero. When the SNR is less than zero the correlation (due to clean patches representing the masking talker) is still evident but less clear, presumably due to the target-specific aspect of the learnt speech model.

4.4. Compact patch encoding

The raw energy representation of a patch is highly redundant and its density model has a very large number of parameters. For example, a model for 12 by 8 patches in a low frequency channel with 200 mixture components, each with a mixing coefficient, a 96 element mean, and a 96×96 element covariance matrix results in a model with nearly 2 million parameters. One way to reduce this figure is to employ a common encoding technique used in image processing: the 2D discrete cosine transform (DCT). The reduction is obtained by truncating the series of DCT coefficients. Reconstruction experiments suggested that only 25% of the DCT coefficients need to be retained. Further, the use of the DCT can be expected to decorrelate the individual parameters somewhat, making the use of a diagonal covariance model possible. However, using the model selection technique described earlier on DCT-reduced patches, it was found that the optimal number of components rises to around 800, suggesting that features are not fully independent. Nevertheless, a 40-fold reduction in the number of model parameters is achieved. The bottom row of figure 5 depicts the likelihood distributions for the DCT representation. Clearly, the two representations are very similar, in that the same trends are evident.

5. Summary and further work

A system which distinguishes speech from other sources on the basis of brief and spectrally-limited regions could form the basis for speech recognition in noise. A model-driven approach to the detection of clean speech patches was introduced, focusing on patch size and the construction of a patch model. The best patch size to ensure that a single source dominates was found to vary with frequency, with brief but spectrally-extensive patches at

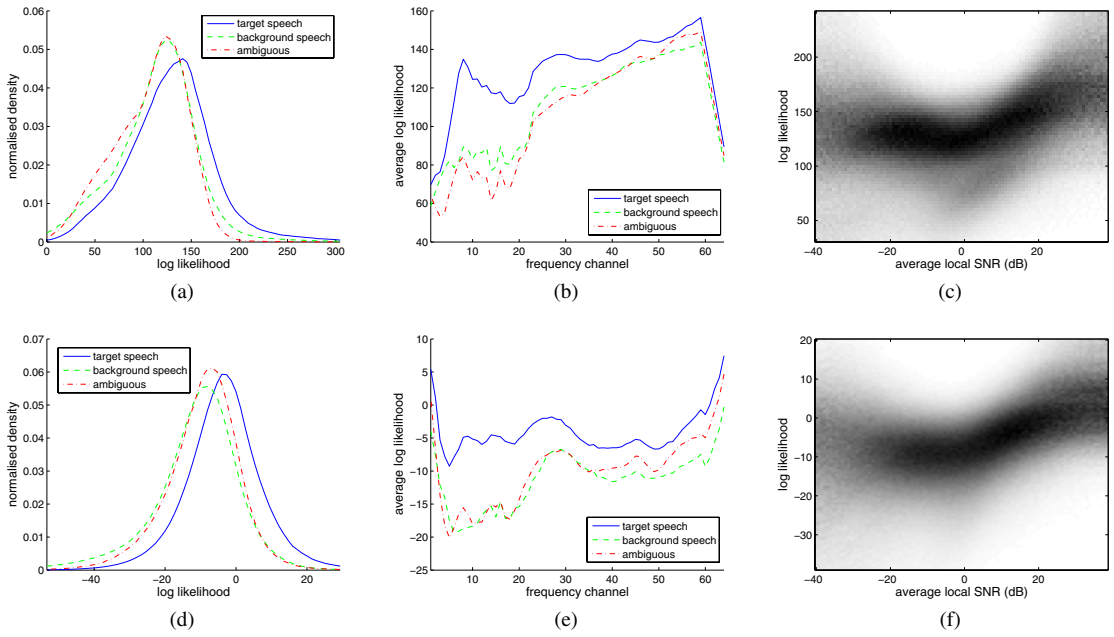

Figure 5: *Top: raw energy (patch size 12 x 8), bottom: DCT (patch size 12 x 8). (a, d) normalised distribution of likelihoods for clean, background and ambiguous pixels in two-talker speech mixtures. (b, e) per-channel average log likelihoods. (c, f) relationship between average local SNR of a patch and its clean speech likelihood.*

higher frequencies and longer but spectrally-narrow patches at low frequencies. Gaussian mixture models for each frequency band were trained on patches of the raw log energy and a more compact representation based on a truncated 2D discrete cosine transform. Both showed similar ability to distinguish the target speaker from the background and from regions dominated by more than a single source.

Future work will investigate "data-driven" methods for feature reduction such as mixtures of probabilistic principal component analysis [15] and seek methods for converting the patch likelihood map into groups which can be recognised by a speech fragment decoder.

6. Acknowledgements

Jonathan Laidler's doctoral studies were supported by the University of Sheffield and by the EU FP6 PASCAL Network of Excellence.

7. References

[1] P. F. Assmann and Q. Summerfield, "The perception of speech under adverse acoustic conditions," in *Speech Processing in the Auditory System*, ser. Springer Handbook of Auditory Research, S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay, Eds. Springer, 2004, vol. 18.

[2] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Communication*, in press, 2007.

[3] A. Bregman, *Auditory scene analysis: the perceptual organization of sound*. The MIT Press, 1990.

[4] M. P. Cooke, *Modelling auditory processing and organisation*. Cambridge University Press, Cambridge, MA, 1993.

[5] G. Brown and M. Cooke, "Computational auditory scene analysis," *Speech Communication*, vol. 8, pp. 351–366, 1994.

[6] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications, neural networks," *Neural Networks*, vol. 13, no. 4–5, pp. 411–430, 2000.

[7] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," *International Conference on Acoustics, Speech and Signal Processing*, vol. '90, pp. 845–848, 1990.

[8] M. J. F. Gales and S. J. Young, "HMM recognition in noise using parallel model combination," *Eurospeech*, vol. 2, pp. 837–840, 1993.

[9] M. Cooke, "A glimpsing model of speech perception in noise," *Journal of the Acoustical Society of America*, vol. 119, pp. 1562–1573, 2006.

[10] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, vol. 45, pp. 5–25, 2005.

[11] M. P. Cooke, M. L. G. Lecumberri, and J. Barker, "The foreign language cocktail party problem: energetic and informational masking effects in non-native speech perception," *submitted to JASA*, 2007.

[12] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, pp. 2421–2424, 2006.

[13] B. C. J. Moore, B. R. Glasberg, C. J. Plack, and A. K. Biswas, "The shape of the ear's temporal window," *Journal of the Acoustical Society of America*, vol. 83, no. 3, pp. 1102–1116, 1988.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. B, no. 39, pp. 1–38, 1977.

[15] M. I. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society*, vol. Series B, no. 61, pp. 611–622, 1999.