

Multi-Resolution Auditory Scene Analysis: Robust Speech Recognition Using Pattern-Matching from a Noisy Signal

Sue Harding¹, Georg Meyer²

¹Department of Computer Science, University of Sheffield, UK

s.harding@dcs.shef.ac.uk

²Department of Psychology, University of Liverpool, UK

g.meyer@liverpool.ac.uk

Abstract

Unlike automatic speech recognition systems, humans can understand speech when other competing sounds are present. Although the theory of auditory scene analysis (ASA) may help to explain this ability, some perceptual experiments show fusion of the speech signal under circumstances in which ASA principles might be expected to cause segregation. We propose a model of multi-resolution ASA that uses both high- and low-resolution representations of the auditory signal in parallel in order to resolve this conflict. The use of parallel representations reduces variability for pattern-matching while retaining the ability to identify and segregate low-level features of the signal. An important feature of the model is the assumption that features of the auditory signal are fused together unless there is good reason to segregate them. Speech is recognised by matching the low-resolution representation to previously learned speech templates without prior segregation of the signal into separate perceptual streams; this contrasts with the approach generally used by computational models of ASA. We describe an implementation of the multi-resolution model, using hidden Markov models, that illustrates the feasibility of this approach and achieves much higher identification performance than standard techniques used for computer recognition of speech mixed with other sounds.

1. Introduction

Despite many decades of study and a large body of experimental evidence, it is still not clear exactly how human listeners deal with the wide variability in acoustic forms that are interpreted as having the same meaning [1], nor how we are able to understand speech with ease in the presence of other sounds [2]. Many theories and models of speech perception have been proposed to explain the first of these abilities e.g. [3, 4, 5]. However, none of these models is widely accepted, and they take little account of the problem of recognising speech in a noisy environment. An approach that deals explicitly with the segregation of speech or other sources from a background of sounds is that of auditory scene analysis (ASA) [6], describing processes by which the auditory signal is segregated into perceptual streams, corresponding to individual sources, using simple grouping and segregation principles based on primitive cues such as frequency and intensity; in this way, a speech stream may be segregated from other sounds. Schema-based processes are responsible for providing a higher-level interpretation of the signal. ASA has been applied to speech and implemented in computational models of auditory scene analysis (CASA) [7]. However, a number of perceptual experiments have shown that listeners can under-

stand speech under conditions that, according to ASA principles, might be expected to cause the signal to segregate into streams that individually are not sufficient to produce a speech percept. For example, listeners are able to perceive a stimulus as a single speech sound even when parts of the signal have different fundamental frequency [8, 9] or are presented to different ears [10]. Thus, although the theory of auditory scene analysis might provide an explanation of how parts of the auditory signal can be segregated into separate streams, its use as a framework for speech perception has been challenged.

We propose a model of multi-resolution auditory scene analysis that aims to resolve the conflict between perceptual experiments showing fusion of the speech signal and ASA principles that would be expected to segregate the signal. The model uses parallel high- and low-resolution representations of the auditory signal, allowing the effects of variability in the speech signal to be reduced while retaining the ability to identify and segregate low-level features of the signal. An important feature of the model is the assumption that features of the auditory signal are fused together unless there is good reason to segregate them. In the model, speech is recognised by matching the signal to previously learned templates without prior segregation of the signal into separate perceptual streams; this contrasts with the approach generally used by CASA models. In cases where the signal matches multiple templates, for example when two voices are present, or when an important cue is obscured by another sound, segregation cues such as fundamental frequency and location will be used to differentiate parts of the signal produced by different sources. However, weak segregation cues may be ignored if they would otherwise prevent the signal being perceived as speech. Only when particularly strong segregation cues are present will primitive segregation processes override the speech schema. Pitch is assumed not to be crucial for speech pattern-matching, although it can aid segregation; it is also used in talker identification and hence speech template selection, as well as for semantic processing via intonation.

An implementation of the multi-resolution model is described below, illustrating the feasibility of this approach and achieving much higher identification performance than standard techniques used for computer recognition of speech mixed with other sounds. The model is described in more detail in [11].

2. Model implementation

2.1. Introduction

The implementation described below tested the feasibility of a number of features of the multi-resolution model, including the

development of pattern templates, the use of a low-resolution representation, and the assumption that fusion is the default, allowing the templates to be matched to the whole signal rather than to segregated features.

Pattern templates were developed using hidden Markov models (HMMs). HMMs can be used to generate templates from examples of speech, using very little prior knowledge of the structure of the data. The HMM templates formed the low-resolution representation within the multi-resolution model. Speech was extracted from a mixed signal using these templates, overriding low-level segregation cues provided by pitch differences.

2.2. Stimuli

Natural vowel-nasal syllables were used for testing the model. A minimum of 40 utterances of each of the six syllables /em/, /en/, /om/, /on/, /um/ and /un/ were collected from 12 talkers (6 male, 6 female). Syllables from one of the talkers were not sufficiently distinct to be consistently identified by human listeners and were excluded from the experiments.

The clean speech syllables were mixed with three non-speech sounds (broadband white noise, a steady sine-wave tone and a rising sine-wave tone), and with three randomly selected syllables produced by other talkers. The mixed stimuli were used to test the ability of model to segregate speech from a mixture of sounds.

2.3. General methods

The vowel-nasal stimuli, sampled at 22050 Hz, were analysed using a Mel-scale filterbank with 60 filters ranging from 0 to 10 kHz and an analysis window of length 10 ms, shifted by 5 ms. This resulted in a fairly high-resolution representation of each syllable consisting of one frame per 5 ms divided into 60 frequency channels. This representation was smoothed to produce a low-resolution representation, using a two-dimensional Gaussian filter specified by the standard deviations σ_t and σ_f in the time and frequency dimensions, respectively; these two parameters could be varied independently. A range of 9 smoothing filters was used to investigate the effects of differing resolution on recognition performance.

The HTK toolkit [12] was used to train speaker-dependent HMMs using half the smoothed clean syllables for each talker, resulting in one template representing each phoneme, for each talker and filter combination. Each template was defined by up to 5 states that could be mapped onto a stimulus. The spectral variation of the stimuli was specified by a mean and variance for each frequency channel in each state of the template. These templates were used to segregate speech from a mixed signal by identifying the state corresponding to each time frame of the test stimulus and using the means and variances to determine how much of the signal was likely to be speech. For each time frame and frequency channel, the signal was compared with a boundary defined by the mean plus the standard deviation (i.e. the square root of the variance) multiplied by a tolerance factor that allowed more or less of the signal to be considered as matching the template. Any excess lying above the boundary was considered not to be speech (Figure 1).

The success of the segregation process in separating the speech from other features of the signal was evaluated by passing the extracted speech back into the recognition process as test data, and calculating the percentage correct identification. The results of this process were also compared with those obtained from testing the mixed signal (i.e. before segregation) against

the HMMs trained on clean speech.

The segregation process described above required the identification of both the correct template and the segment boundaries in the test stimulus before each time frame could be matched to a state of an HMM. Determining these properties for speech in noisy conditions using HMMs is difficult and is a major problem in automatic speech recognition.

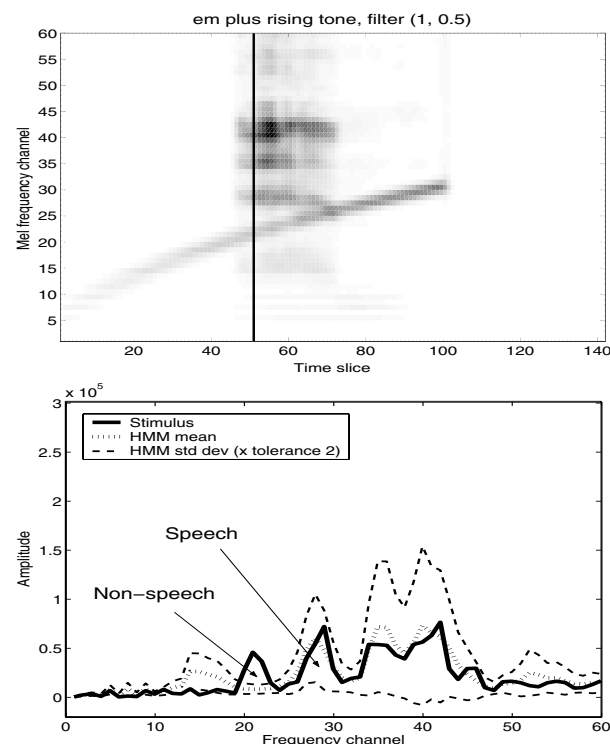


Figure 1: Upper part of the figure: the syllable /em/ mixed with a rising pure tone mixed at 0 dB SNR. Lower part: a cross section through the stimulus, along the vertical line shown above, showing the amplitudes in each frequency channel for a single time frame (solid line). The means (dotted line) and standard deviations multiplied by the tolerance factor (dashed line) are shown for the state of an HMM that corresponds to the stimulus time frame. The peak corresponding to the pure tone extends above the tolerance boundary. A tolerance factor of 2 was used in this case.

The experiments described below used clean segmentation information, obtained from testing the clean speech stimuli against HMMs that were trained on clean speech. The segmentation for each stimulus was then used during the process of segregating that stimulus from its mixture with other sounds. This is obviously a short-term solution to the problem of finding the correct segmentation, discussed further below.

Two experiments were performed: the first tested the effect of changing the amount of smoothing, and the second the effect of signal-to-noise ratio (SNR) on recognition performance. Both experiments used an intermediate tolerance factor of 2.

3. Experiment 1: smoothing filter

Experiment 1 tested the effect of different amounts of smoothing on the recognition performance of the model. Three forms of the signal were available for testing recognition performance:

the original clean speech; the clean speech mixed with other sounds; and the speech after extraction from the mixture.

3.1. Stimuli

Nine smoothing filters (including the unsmoothed case) were used and were applied to clean, mixed and extracted syllables, where the mixed syllables were either speech/non-speech or mixed talkers. Speech was extracted from the mixtures using the process described in section 2.3. Each type of signal (clean, mixed and extracted) was tested against the HMM templates derived from clean speech smoothed by the same filter. For the mixed talkers, the two largest filters were omitted from the experiment once it was clear that a similar pattern was emerging as for the speech/non-speech stimuli.

3.2. Results

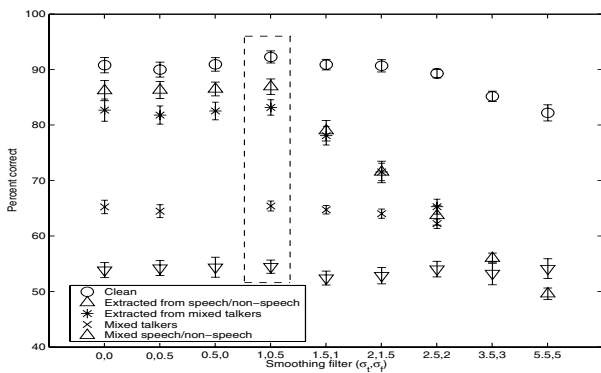


Figure 2: Effect of smoothing filters on the recognition of clean speech, mixed speech/non-speech, mixed talkers, and speech extracted from these two mixtures. Error bars show standard errors. Results for the best filter are highlighted.

Figure 2 shows the mean performance for each filter over all talkers for clean, mixed speech/non-speech, mixed talkers, and speech extracted from these two mixtures. A significant effect of filter on recognition performance was found, using a one-way ANOVA, for the clean speech ($F(1,8) = 7.94$, $p < 0.01$) and for the speech extracted from the speech/non-speech mixture and mixed talkers ($F(1,8)=88.50$, $p < 0.01$ and $F(1,6)=18.31$, $p < 0.01$, respectively), but not for the mixed speech/non-speech ($F(1,8)=0.2198$, $p=0.987$), nor for the mixed talkers ($F(1,5)=1.4234$, $p=0.229$).

Recognition performance, using the best smoothing filter (1,0.5), for speech extracted from a speech/non-speech mixture (86.9 %) or extracted from mixed talkers (83.1 %), was approaching that for clean speech (92.3 %); the performance was much better than that for the mixed signals prior to segregation (54.5 % and 65.4 % respectively), showing that the segregation process was successful. Smoothing up to filter (1,0.5) caused no deterioration in recognition performance, but performance was reduced for larger filters.

4. Experiment 2: signal-to-noise ratio

Automatic speech recognisers using HMMs normally achieve a much lower recognition performance for low signal-to-noise ratio (SNR) compared with human performance. Experiment 2 tested the effect of different SNR on the recognition perfor-

mance for the mixed and extracted speech.

4.1. Stimuli

For experiment 2, a single smoothing filter (1,0.5) was used, and the mixed speech/non-speech and mixed talker stimuli were mixed at a range of signal-to-noise ratios, namely -24, -18, -12, -6, 0, 6, 12 and 18 dB. The mixed and extracted speech was tested against the clean speech templates as in the previous experiment.

4.2. Results

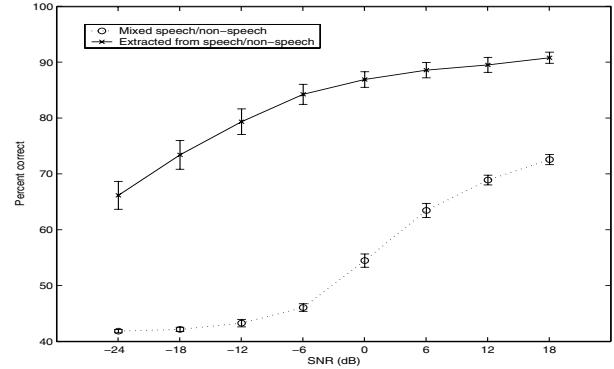


Figure 3: Effect of signal-to-noise ratio on the recognition of speech mixed with non-speech, or extracted from this mixture, for a single smoothing filter (1,0.5). Error bars show standard errors.

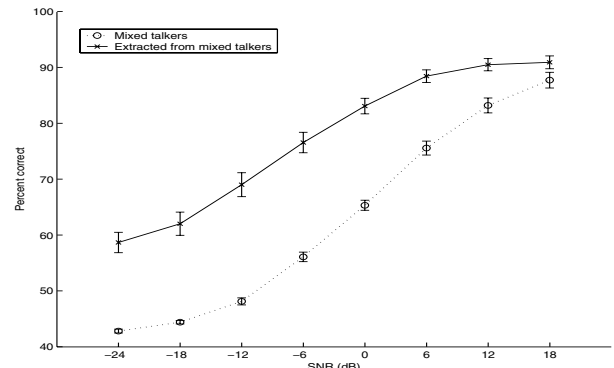


Figure 4: Effect of signal-to-noise ratio on the recognition of speech mixed with other speech, or extracted from this mixture, for a single smoothing filter (1, 0.5). Error bars show standard errors.

The recognition performance before and after segregation for mixed speech/non-speech and mixed talkers is shown in figures 3 and 4. In each case, recognition performance was much higher for the extracted speech than for the mixed speech, as in experiment 1, and showed improvement with increasing SNR.

In order to compare the effect of SNR on the two recognition methods (i.e. mixed data before segregation and extracted data after segregation), each of the four sets of data was fitted with a Boltzmann curve $y = \frac{A_1 - A_2}{1 + e^{(x - x_0)/dx}} + A_2$. The lower and upper asymptotes (A_1 , A_2) and the slope (dx) were constrained to be identical for each of the two types of data (mixed

speech/non-speech or mixed talkers), since the lower asymptote was expected to be around chance performance (41.7 %) and the upper asymptote close to the performance for clean speech (92.3 % in experiment 1). Fitting the curves to the mixed/non-speech data gave $A_1 = 38.1 \pm 1.8$, $A_2 = 91.3 \pm 1.2$, $dx = 10.7 \pm 1.2$, $\chi^2 = 2.21$ and $R^2 = 0.995$, with x_0 (mixed) = 9.2 ± 1.0 and x_0 (extracted) = -25.3 ± 1.3 , i.e. an improved performance equivalent to an increase of 34.5 dB for the extracted speech compared with the mixed speech/non-speech stimuli. For the mixed talker data, the fitting parameters were: $A_1 = 37.6 \pm 4.3$, $A_2 = 94.0 \pm 3.6$, $dx = 9.9 \pm 2.0$, $\chi^2 = 4.68$ and $R^2 = 0.988$, with x_0 (mixed) = 0.1 ± 1.7 and x_0 (extracted) = -15.5 ± 1.7 , i.e. an improved performance of 15.6 dB for the extracted speech compared with the mixed talker stimuli. Thus a decrease in SNR had a much smaller effect on the extracted speech than on the mixed speech.

5. Discussion

Smoothing the signal by filter (1,0.5) or smaller (equivalent to 30 ms and 3 Mel frequency channels) caused no reduction in performance, supporting a basic premise of the multi-resolution model: that speech pattern templates, sufficient for identifying phonemes, can be developed from smoothed representations of examples of speech without detailed knowledge of the low-level features of the speech signal. Recognition performance for the extracted speech was close to that for clean speech, showing the success of the segregation process.

A major assumption of the multi-resolution model is that the acoustic signal is fused into a whole unless there is good reason to segregate parts of it. In this implementation, no attempt was made to identify cues for segregation, and the whole signal was matched to the speech templates. Although parts of the signal corresponding to the non-speech or secondary talker were included in the extracted portion of the signal, this caused only a small reduction in recognition performance for the extracted speech compared with clean speech.

A factor which would be expected to reduce the performance in the model is the use of a more realistic segmentation method. In the implementation described above, clean segmentation was used, obtained from testing the clean speech stimuli against the HMM templates. An alternative segmentation method would be expected to produce better performance than that found by testing mixed signals against the HMMs, although the performance is unlikely to reach that obtained using the clean segmentation. Such a reduction in performance might be offset if pitch differences were taken into account, since those elements with very different characteristics to the speech would be more likely to be segregated and might affect the best template match. Pitch information would also be useful for determining segmentation.

Automatic speech recognisers do not perform well in noisy environments, particularly when the noise is not easy to predict. As signal-to-noise ratio decreases, the performance of automatic speech recognisers deteriorates much more rapidly than human recognition performance. For the multi-resolution model, recognition performance deteriorated with decreasing SNR, but the deterioration was much less marked for the extracted speech than for the mixed speech. For the mixed speech/non-speech, the segregation process produced an improvement in performance over the mixed signal equivalent to an increase in SNR of 34.5 dB; for the mixed talkers the improvement was 15.6 dB. These improvements are very large and are consistent with the behaviour seen for human listeners [2].

Overall, the results obtained from this implementation support the proposals for the multi-resolution model. Speech can be identified using low-resolution pattern templates developed from examples of speech without the need for identification of low-level acoustic features, and may be segregated from other sounds by matching the whole acoustic signal to previously learned templates. Extracting speech in this way produced good recognition performance, even without using pitch information to segregate features of the signal that were clearly from different sources. Recognition performance for the segregated speech was close to that obtained for clean speech, and much higher (i.e. equivalent to an increase in SNR of more than 15 dB) than that obtained when attempting to recognise the speech using the whole signal.

6. Acknowledgements

SMH was funded by EPSRC research studentship no. 99304828.

7. References

- [1] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.*, vol. 24, pp. 175–184, 1952.
- [2] G. A. Miller, G. A. Heise, and W. Lichten, "The intelligibility of speech as a function of the context of test materials," *J. Exp. Psychol.*, vol. 41, pp. 329–335, 1951.
- [3] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *J. Acoust. Soc. Am.*, vol. 111, pp. 1872–1891, Apr 2002.
- [4] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, pp. 532–556, 1989.
- [5] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, pp. 1–36, 1985.
- [6] A. S. Bregman, *Auditory Scene Analysis: the perceptual organization of sound*. Cambridge, MA: MIT Press, 1990.
- [7] D. F. Rosenthal and H. G. Okuno, eds., *Computational auditory scene analysis*. London: Erlbaum, 1997.
- [8] C. J. Darwin, "Perceptual grouping of speech components differing in fundamental frequency and onset-time," *Q. J. Exp. Psychol.*, vol. 33, pp. 185–207, 1981.
- [9] J. F. Culling and C. J. Darwin, "Perceptual separation of simultaneous vowels: Within and across formant grouping by F0," *J. Acoust. Soc. Am.*, vol. 93, pp. 3454–3467, June 1993.
- [10] D. E. Broadbent and P. Ladefoged, "On the fusion of sounds reaching different sense organs," *J. Acoust. Soc. Am.*, vol. 29, pp. 708–710, 1957.
- [11] S. M. Harding, *Multi-resolution auditory scene analysis for speech perception: experimental evidence and a model*. PhD thesis, Keele University, submitted 2003.
- [12] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book (for HTK version 3.0)*. Available from <http://htk.eng.cam.ac.uk/>: Microsoft Corporation, 2000.