ForumAcusticum
2 0 0 5

Towards Solving the Cocktail Party Problem through Primitive Grouping and Model Combination

Jon Barker, André Coy
Department of Computer Science, University of Sheffield,
211 Portobello Street, Sheffield, S1 4DP, United Kingdom, e-mail: **{j.barker,a.coy}@dcs.shef.ac.uk**

This paper considers the problem of recognising speech in the presence of a competing speaker in a single channel condition. A solution is suggested that integrates the top-down aspect of model combination approaches, with the bottom-up aspect of the primitive grouping processes proposed by Computational Auditory Scene Analysis models. First, primitive grouping processes locate local spectro-temporal fragments in which all spectro-temporal points are dominated by a common source. Each fragment may belong to either the foreground source or the background source. Using HMM-based models of both sources, the top-down processes jointly search for the best fragment labeling and model state sequences. The present paper uses simulated grouping processes to demonstrate how primitive grouping can potentially boost recognition performance above that of a purely top-down system. The system is tested using simultaneous continuous digit sequences mixed at 0 dB SNR. In mixed gender conditions, with sufficient primitive grouping, the target speech source can be reliably recognised, even when using only a single Gaussian to model the competing speech source.

1 Introduction

Speech recognition technology has been under development for over 30 years, but despite great progress there remain key problems that prevent the technology being deployed more widely. Chief among these problems is a lack of robustness to additive noise. Although many solutions have been pursued, most techniques do not work well when the noise source is non-stationary, and perform particularly poorly when the noise is a competing speaker (the so-called 'cocktail party' problem [1]).

One approach to the cocktail party problem is to use two or more microphones, and decompose the acoustic mixtures arriving at each by using blind source separation techniques, typically based on independent components analysis (ICA) [2]. Such microphone-array techniques can work well in certain circumstances, but require at least as many microphones as there are acoustic sources, and say nothing about how listeners are able to understand complex acoustic mixtures, even in situations where only one channel is available.

So how do we understand speech mixtures in single-channel conditions? There exist two contrasting approaches, that are often characterised as being 'top-down' or 'bottom-up'. Top-down approaches attempt to explain the acoustic mixture by combining models of individual acoustic sources. They rely on techniques for searching across the combined model space to find the best match to the observed data. One implementation of this idea is the factorial HMM (FHMM) (e.g. [3]), which forms a composite HMM through a factorial combination of the state-spaces of the individual source models. These techniques suffer from two key problems. First, there is combinatorial explosion in the size of the state space of the composite HMM that can render the approach computationally intractable. Second, problems with the standard HMM (e.g. poor duration modelling) become compounded when models are combined leading to a fatal lack of constraint in the composite model.

In contrast to model-driven approaches, 'bottom-up' approaches work by trying to segregate the acoustic sources using 'primitive' properties that are common to all acoustic sources. These approaches have been inspired by the Auditory Scene Analysis account of auditory perception [4]. Typically, they act on a spectro-temporal representation, and work by grouping elements that appear to belong to the same source. As they exploit properties that are common to all sound sources they do not rely on specific models of individual sources. Although bottom-up processes can perform a partial analysis of an acoustic mixture, they are seldom sufficient to form an unambiguous interpretation. In general, there may be strong evidence for grouping elements to form local spectro-temporal 'fragments', but there may be few reliable cues for grouping local fragments over longer time scales.

Although both 'top-down' and 'bottom-up' solutions are incomplete, their strengths are possibly complementary. This paper explores a model combination approach which incorporates bottom-up constraint. The system is tested using a simultaneous connected digit sequence task. The paper first presents the standard factorial HMM-based model combination approach. Section 3 motivates the integration of primitive grouping processes into the model combination approach, and Section 4 details one possible way this may be achieved. Section 5 describes a series of experiments comparing model combination either with or

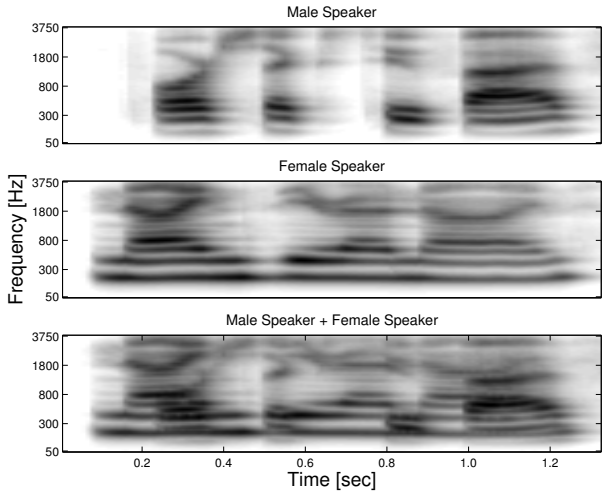

Figure 1: A comparison of 'auditory spectrograms' of a male utterance (top), a female utterance (centre) and of an equal energy mixture of the two utterances (bottom).

without grouping constraints. The results of these experiments are discussed in Section 6.

2 The max approximation

Figure 1 shows a log-energy spectro-temporal representations of a male speaker uttering a digit sequence (top panel), a female speaker uttering a different digit sequence (middle panel), and of the result of mixing the two speakers such that they have equal average energy (bottom panel). Note that it is possible to visually segregate the mixture into regions that appear to have come from the male spectrogram and regions that appear to have come from the female spectrogram. This is a consequence of the max approximation, which observes that if the signal $y(t)$ is the addition of two source $x(t)$ and $z(t)$, then in the frequency domain,

$$log|Y(f)| \approx max(log|X(f)|, log|Z(f)|) \qquad (1)$$

The error in the approximation is greatest when the signals have equal energy at a particular frequency (see Figure 2). For speech signals, which have a time-varying and 'peaky' spectrum, most frequency channels will be dominated by one or other of the sources and (1) remains a good approximation. In Figure 2 the approximation error is shown, and it can be seen that it is negligible in most channels.

Model combination approaches can exploit the max approximation to simplify the statistical modelling of the mixed signal. Consider isolated sources, $X(f)$ and $Z(f)$, and their mixture, $Y(f)$. If the sources are assumed to be independent then under the max approximation,

$$F_{\mathbf{y}}(\lambda) = F_{\mathbf{x}}(\lambda)F_{\mathbf{z}}(\lambda) \qquad (2)$$

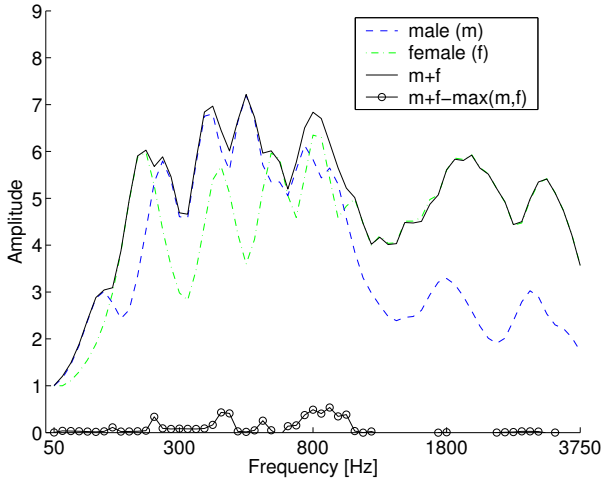

Figure 2: Compressed spectral profiles taken at frame 23 of Figure 1. When acoustic sources are added the compressed spectrum of the result can be approximated by taking the maximum of the compressed spectra of the unmixed sources. The error in the approximation, plotted with the 'o' symbol, is negligible for most channels.

where $F_{\mathbf{y}} = P(\mathbf{y} < \lambda)$. The PDF of $\mathbf{y}$ is obtained by differentiating (2) with respect to λ,

$$p_{\mathbf{y}}(\lambda) = p_{\mathbf{x}}(\lambda)F_{\mathbf{z}}(\lambda) + p_{\mathbf{y}}(\lambda)F_{\mathbf{x}}(\lambda) \qquad (3)$$

This result was noted by Nadas et al. [5], but was first used much earlier by Varga and Moore [3] in their presentation of a factorial HMM approach to robust ASR. In the case where the PDFs of $\mathbf{x}$ and $\mathbf{z}$ are modelled using a Gaussian mixture model with mixtures having diagonal covariance matrices, expressions for the CDF terms, $F_{\mathbf{z}}$ and $F_{\mathbf{x}}$, in (3) can be easily derived [6].

If both sources are modelled using HMMs, then the combined source can be modelled using an FHMM in which the emission probabilities of the combined states are formed from the emission probabilities of the independent states according to (3) – see, for example, [6]. In the experiments reported in Section 5, the masking source, $z(t)$, has been modelled using a single distribution, $p_{\mathbf{y}}(\lambda)$. In this case we simply take the HMM for the the target source, $x(t)$, and replace the emission PDFs in each state $p_{\mathbf{x},q}(\lambda)$ with the emission PDF $p_{\mathbf{y},q}(\lambda)$ given by (3). Results obtained using this model are presented in Section 6.

3 Grouping constraints

A problem with the model combination approach is that for any given frame of data there may exist many different valid interpretations. This problem is illustrated in Figure 3. On the left is shown an observation of the spectrum of the mixture, y, with two peaks, which may, for

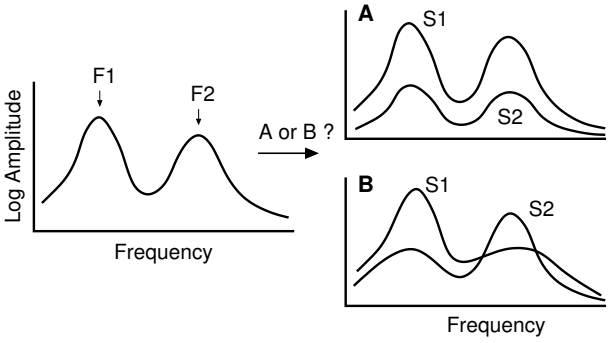

Figure 3: The observed spectral profile on the left can be decomposed into two sources in many different ways. Two possible contrasting explanations are shown on the right. These explanations may or may not be consistent with primitive grouping constraints.

example, correspond to speech formants. On the right there are shown two alternative explanations of this observation: in the first case, **A**, both peaks are contributed by source **S1**, and source **S2** having lower energy is not contributing to the the observation; in the second case, **B**, the lower frequency peak is contributed by source **S1**, and the higher frequency peak by source **S2**. The likelihood of these two interpretations of the observation may be quite similar.

Now consider that there exists some independent evidence that suggests that the energy in the peaks at F1 and F2 is being contributed by *different* sources. For example, this evidence may come from a binaural mechanism which estimated that the energy in the two frequency regions was arriving from different directions; or it may come from an analysis of the temporal fine structure in these frequency regions that reveals that the signals in channels around F1 and F2 have different periodicities; or it may come from an analysis of the spectrogram over longer time windows that reveals that the format peak F1 onset at a different time to formant peak F2. (Generally, such evidence may arise from a CASA-style primitive grouping front-end applied to the signal). Evidence suggesting F1 and F2 were due to different sources, would clearly mediate against interpretation A in Figure 3, i.e. bottom-up analysis may constrain the interpretations considered by a purely top-down system.

4 The fragment decoding model

In this paper the value of grouping constraints is demonstrated using a version of the 'speech fragment decoding' approach proposed by Barker et al. [7] that has been extended to incorporate a model of the background noise source.

The standard ASR framework finds the acoustic state sequence, $Q = \{q_1, q_2, ..., q_n\}$, with the highest posterior probability given observations of the speech acoustics, X, using,

$$Q_x = \underset{Q_x}{\operatorname{argmax}}\ p(Q_x|X) = \underset{Q_x}{\operatorname{argmax}}\ p(X|Q_x)p(Q_x) \tag{4}$$

The fragment decoding approach extends the usual ASR search for the best acoustic state sequence, to simultaneously search for the best state sequence *and* best foreground/background segmentations given the mixed acoustic signal, Y. Here we further extend the search to consider the state spaces of two independent sources, Q_x and Q_z,

$$Q_x, Q_z, S = \underset{Q_x, Q_z, S}{\operatorname{argmax}}\ p(Q_x, Q_z, S|Y) \tag{5}$$

$$= \underset{Q_x, Q_z, S}{\operatorname{argmax}}\ p(Q_x, Q_z|S, Y)P(S|Y) \tag{6}$$

The segmentation, S, can be considered as a binary mask which labels each spectro-temporal point as either being dominated by the foreground source, or by the background source. Under the max approximation, being given the segmentation means knowing, for both q_x and q_z, the identity of the channels in which the model is generating the observed energy and those in which it is producing energy less than that observed. If this is known then the probability of the state of one source is entirely independent of that of the other,

$$p(Q_x, Q_z|S, Y) = p(Q_x|S, Y)p(Q_z|S, Y) \tag{7}$$

Using (7) and Bayes' theorem, (5) can be rewritten as,

$$Q_x, Q_z, S = \underset{Q_x, Q_z, S}{\operatorname{argmax}}\ p(Y|Q_x, S)p(Y|Q_z, S)p(Q_x)p(Q_z)P(S|Y) \tag{8}$$

The terms $p(Y|Q_x, S)$ and $p(Y|Q_z, S)$ are approximated as a product over independent frames in the standard fashion, e.g. ,

$$p(Y|Q_x, S) = \prod_{i=1}^{n} p(y_i|q_{xi}, s_i) \tag{9}$$

where s_i is the segmentation of an individual frame. At each frame $p(y_i|q_{xi}, s_i)$ and $p(y_i|q_{zi}, s_i)$ are estimated using a 'missing data model' [8].

The term $P(S|Y)$, known as the 'Segmentation Model,' models the bottom-up processes. Given the data some segmentations are more probable than others. For example, in a binaural system, a segmentation would have low probability if energy regions which have been labelled as part of the foreground do not appear to emanate from the same position.

The experiments reported here start by assuming that in each frame, f_t, the frequency channels can be perfectly

Figure 4: Examples of the fragments used to model the segmentation of one particular utterance pair. Fragments are defined over a consecutive window of frames. From top to bottom, the figure shows fragments spanning windows of 1 frame, 10 frames, 40 frames, and all frames. Each fragment is shown in a different shade of grey.

segregated into two sets, s_t and s'_t, describing the channels dominated by each of the two sources present. If this processing is applied to each of N frames, then 2^N valid segmentations can be formed by taking each frame and choosing either s_t or s'_t to be the foreground source, i.e. in each frame there are two complementary 'fragments' and the segmentation is built up by choosing a single fragment from each frame. This condition is illustrated in the top panel of Figure 4. In a more constrained condition, the segregation of each frame can be applied over longer time windows to form source fragments that extend over several frames. This is shown for 10 and 40 frame windows in the two centre panels of Figure 4. At the extreme the whole utterance can be perfectly segmented to form two fragments (bottom panel of Figure 4). For each condition, if there are M segregated windows, there are 2^M valid segmentations of the entire utterance. When the windows are short this may be an impractically large number over which to search. An efficient algorithm which can evaluate the entire search space without the need for pruning is presented in [7].

5 Experiments

5.1 Test data and model training

The test data was constructed from the 1001 utterances of test set A of the Aurora 2 connected digit sequence corpus [9]. An end-point detection algorithm was used to remove initial and final silence [10]. The end-pointed utterances were then ordered by length and each signal was paired with its neighbour to create 1000 utterance pairs. Pairs in which either utterance contained less than 3 digits were discarded. Pairs in which both utterances started with the same digit were also discarded – this allowed the first digit to act as a unique identifier for the target utterance. This processing resulted in 262 male-female utterance pairs which formed the mixed-gender test set, and 283 male-male utterance pairs to form the matched-gender test set. The acoustic mixtures were constructed by adding the signal pairs in the time domain. The shorter of each pair was padded with zeros (equally at either end) to match the size of the longer signal.

Feature vectors were formed by filtering the signals with a 64 channel gammatone filter bank with centre frequencies equally spaced on an ERB scale from 50 Hz to 3850Hz. The instantaneous Hilbert envelope at the output of each filter was smoothed with a 1st order filter (with an 8 ms time constant) and sampled at a frame rate of 100 Hz. Cube root compression was then applied to the envelope values. We refer to this representation as an 'auditory spectrogram'. Note, cube root compression was used rather than log compression, as it is a better model of the compression performed by the ear when converting energy into perceived loudness [11], and it has been shown in the past to produce better recognition results when using missing data techniques [12]. The 'max' approximation remains reasonably valid in the cube-root domain (see Figure 2).

To model the target speaker, whole word gender dependent HMMs were trained using the 4220 utterances of each gender in the Aurora clean training set, and the same 'auditory spectrogram' representation. Each HMM had 16-states in a straight-through topology. Each state was modelled with a mixture of 7 Gaussian distributions each with a diagonal covariance matrix. A single-state silence model was constructed to model inter-digit pauses. These models have been used previously in [12, 13]. The masking speaker was modelled using either a single Gaussian distribution with diagonal covariance, or a mixture of 5 Gaussian distributions with each mixture having a diagonal covariance matrix. Male and female versions of the masking speaker models were trained.

5.2 Simulated grouping constraints

Primitive grouping was simulated by using knowledge of the signals prior to mixing. Spectro-temporal fragments like those shown in Figure 4 were constructed as follows. A binary segmentation, $B(f,t)$, of the mixed spectrogram $S_y(f,t)$ was constructed by labelling each spectro-temporal point as either 0 or 1 depending on which source is dominant,

$$B(f,t) = \begin{cases} 1 & : \quad S_x(f,t) > S_z(f,t) \\ 0 & : \quad S_x(f,t) \leq S_z(f,t) \end{cases} \qquad (10)$$

From this binary segmentation a set of fragments, of duration w frames, can be defined by assigning each spectro-temporal point a fragment label, $F(f,t)$,

$$F(f,t) = B(f,t) + 2 * floor\left(\frac{t}{w}\right) \qquad (11)$$

where $floor(x)$ is the largest integer value less than x .

The simulated bottom-up grouping defines a set of fragments, but it is not know which fragments belong to which source. At one extreme, when w equals 1, the top down search has to make a fresh decision at every frame (Figure 4, top panel). At the other, when w is set equal to the number of frames in the utterance, there exist only two fragments (Figure 4, bottom panel), and the top down search simply has to identify which of these is the foreground source and which the background.

5.3 Recognition experiments

In all experiments the task was to recognise the male speaker that was uttering the digit sequence that started with a given digit. This was modelled by supplying a grammar which specified the first digit then allowed an arbitrary length sequence of arbitrary digits to follow. Recognition accuracies were computed over the unknown digits only.

Baseline results were measured using the purely top-down model described in Section 2. In this case the male digit foreground HMM, was combined with either the single-state female or single-state male model for the 'mixed gender' and 'matched gender' conditions respectively. This baseline was compared with results of the fragment decoding approach using the same models and with varying size fragments.

6 Results and discussion

Results for the mixed-gender case are shown in Figure 5 (top). Using the male HMMs without any modification to the emission distributions produces a recognition accuracy of 30.2%. Introducing the simple Gaussian female model and using the unconstrained top-down approach increases performance substantially to 54.7%. However, if it were known which regions were dominated by the male source it is possible to achieve a result of 94.6% using the 'missing data' techniques described in Cooke et al. [8]. The fragment decoding approach can achieve this result when there are just two large fragments. If the

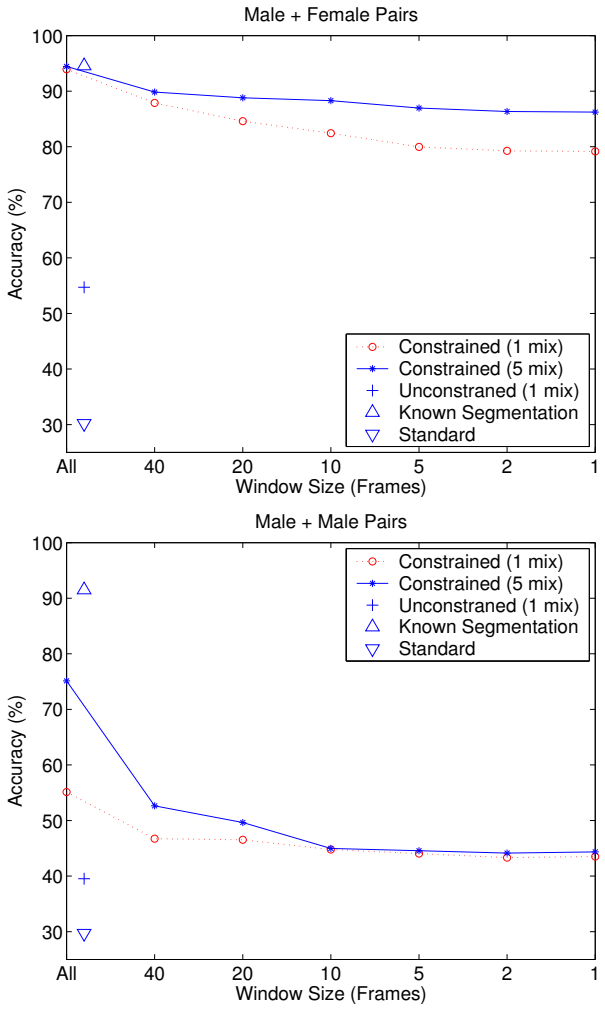

Figure 5: Recognition results using mixed-gender utterances (top) and matched-gender utterances (bottom).

fragment duration is reduced – weakening the bottom-up constraints – performance decreases, but it remains at 79% even for 1-frame fragments. By increasing the detail of the background model by using 5-mixtures rather than a single Gaussian, the performance rises from 79% to 86%.

The point to observe is that the introduction of constraints has allowed the system to perform well despite the crudity of the background model. If very simple background models can be employed then the problems of the factorial explosion in the size of the state-space of the combined model are avoided. The idea that the background is not modelled with the same level of detail as the foreground is also commensurate with the experiences of listeners – whereas FHMM systems which use a full model for both sources are able to simultaneously transcribe both components of a speech mixture, humans can only attend to one speaker at a time.

Results for the matched gender condition show a differ-

Table 1: Recognition accuracies using an HMM trained on clean data (standard), an unconstrained top-down model with a Gaussian noise model (unconstrained), or a missing data approach given a known segmentation (segmented).

	standard	unconstrained	segmented
M + F	30.2	54.7	94.6
M + M	29.7	39.5	91.5

Table 2: Fragment decoding recognition accuracies using either a 1 or 5 mixture noise model, and fragments of width varying from a single frame to the full utterance (see Figure 4).

	frames	1	2	5	10	20	40	all
M+F	1 mix	79	79	80	82	85	88	94
	5 mix	86	86	87	88	89	90	95
M+M	1 mix	44	43	44	45	47	47	55
	5 mix	44	44	45	45	50	53	75

ent pattern. Accuracy for the standard system and for the known segmentation are a little lower than those obtained in the mixed gender condition. The small drop in performance presumably arises as male speech is a better energetic masker of male speech – with a female masker there are glimpses of the target source between the resolved harmonics of the first formant region. The top-down model and the fragment decoding model perform much worse – this is not surprising since both utterances are male there is no way for the models to distinguish them. The fact that the first digit is given means that in theory the utterances could be tracked separately using speaker differences and temporal continuity cues. However, speaker differences are not represented in the speaker-independent HMMs and temporal structure is poorly modelled. In the case where there are only two fragments which span the whole utterance then performance is good – this is because the given-first digit 'attracts' the correct fragment and no further fragment labelling decisions have to be made.

Clearly, since the 'fragment decoding' results reported here employ fragments derived from simulated grouping process they serve merely as a 'proof of concept' for the approach. The results illustrate the idea that bottom up grouping principles can constrain the model space of top-down techniques, and allow for good recognition results in the absence of detailed models of the noise background. Real grouping techniques that can identify fragments without prior knowledge of the unmixed signals are being developed [13]. The next stage will be to test whether the imperfect fragments identified by such techniques can produce similar improvements over the purely top-down baseline system.

References

[1] E.C. Cherry, "Some experiments on the recognition of speech, with one and two ears," *Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] A.J. Bell and T.J. Sejnowski, "An information-maximisation approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1004–1034, 1995.

[3] A.P. Varga and R.K. Moore, "Hidden Markov model decomposition of speech and noise.," in *Proc. ICASSP '90*, Alburquerque, NM, Apr. 1990, pp. 845–848.

[4] A.S. Bregman, *Auditory Scene Analysis*, MIT Press, 1990.

[5] A. Nadas, D. Nahamoo, and M.A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Transactions on Speech and Audio Processing*, vol. 37, no. 10, 1999.

[6] A. Deoras and M. Hasegawa-Johnson, "A factorial HMM approach to simultaneous recognition of isolated digits spoken by multiple talkers on one audio channel," in *Proc. ICASSP '04*, Montreal, Canada, May 2004, vol. 1, pp. 857–860.

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

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

[9] H.G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ICSLP '00*, 2000, vol. 4, pp. 29–32.

[10] L.R. Rabiner and M.R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *The Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975.

[11] H. Fletcher, "Physical measurements of audition and their bearing on the theory of hearing," *Journal of the Franklin Institute*, vol. 196, no. 3, pp. 289–326, 1923.

[12] J.P. Barker, M.P. Cooke, and P.D. Green, "Robust ASR based on clean speech models: An evalutation of missing data techniques for connected digit recognition in noise," in *Proc. Eurospeech '01*, Aalborg, Denmark, 2001.

[13] A. Coy and J. Barker, "Recognising speech in the presence of a competing speaker using a 'speech fragment decoder'," in *Proc. ICASSP '05*, Philadelphia, PA, Mar. 2005.