SPEECH SEPARATION BASED ON THE STATISTICS OF BINAURAL AUDITORY FEATURES

Guy J. Brown, Sue Harding and Jon P. Barker

Department of Computer Science, University of Sheffield, United Kingdom
{g.brown,s.harding,j.barker}@dcs.shef.ac.uk

ABSTRACT

A computational auditory scene analysis (CASA) system is described, in which sound separation according to spatial location is combined with the 'missing data' approach for automatic speech recognition. Time-frequency masks for the missing data recognizer are derived from the statistics of interaural time and level differences; these masks identify acoustic features that constitute reliable evidence of the target speech signal. It is demonstrated that this approach yields good performance in a challenging environment, in which a target voice is contaminated by another talker and reverberation. The ability of the system to generalize to source-receiver configurations that were not encountered during training is discussed.

1. INTRODUCTION

Automatic speech recognition (ASR) remains a challenging problem in noisy and reverberant environments, but human speech recognition performance in such conditions is relatively robust. One factor that might underlie this difference is that human listeners analyze the acoustic input using two ears, whereas ASR systems typically take their input from a single audio channel.

Binaural processing contributes to human hearing in several ways. Firstly, human listeners can localize sounds in space by comparing differences in sound level and time-of-arrival at the two ears. These cues are known as interaural level difference (ILD) and interaural time difference (ITD) respectively. Secondly, binaural mechanisms counteract the effects of reverberation by suppressing echoes. Finally, binaural hearing contributes to the ability of listeners to attend to a target source in the presence of other interfering sounds. Most simply, listeners may attend to the ear in which the signal-to-noise ratio (SNR) is favorable. However, binaural mechanisms that cancel interference or group acoustic energy that originates from the same spatial location also play a role.

This paper describes a computational auditory scene analysis (CASA) system which exploits binaural processing in order to improve the robustness of ASR in multisource, reverberant environments. In the first stage of our system, acoustic features (spectral energies) and binaural features (ILD and ITD) are obtained from an auditory model. The statistics of the binaural features are used to derive a time-frequency (T-F) mask, in which each element indicates whether the corresponding acoustic feature is reliable (dominated by the target sound) or unreliable (dominated by interference). In the second stage, the acoustic features and T-F mask provide the input to a 'missing data' ASR system, which treats reliable and unreliable features differently during decoding.

The current paper extends our previous work [1] in two respects. Firstly, we label T-F regions as unreliable if they have a low interaural coherence; this may be regarded as a model of the precedence effect [2]. Secondly, we investigate the ability of the system to generalize to different configurations of the target source and receiver.

2. METHOD

2.1. Corpus

The input to the system consisted of a target speech signal and a concurrent, but spatially separated, speech masker. Utterances were drawn from the TIDigits corpus, and consisted of a sequence of between one and seven digits spoken by a male talker (possible digits were 'one' to 'nine', 'zero' and 'oh'). All data were sampled at a rate of 20 kHz. The target and masking signals were spatialized and reverberated by the ROOMSIM model of small-room acoustics [3], using a simulated room of size 6 m $\times$ 4 m $\times$ 3 m (length $\times$ width $\times$ height). The receiver was a simulated KEMAR manikin in the center of the room, 2 m above the ground. Target speech was spatialized at an azimuth of 0 degrees, and the masker was placed at azimuths of 5, 7.5, 10, 15, 20, 30 or 40 degrees. Target and masking sources were spatialized at a radial distance of 1.5 m from the center of the head.

All surfaces of the room had reverberation characteristics consistent with 'acoustic plaster', giving a T_{60} reverberation time of 0.34 sec. To investigate the effect of the relative position of the source and receiver, two conditions were considered in which the target-receiver axis was rotated by 5 degrees and 20 degrees about the center of the head relative to the longer wall of the room (Fig. 1).

This work was funded by EPSRC grant GR/R47400/01.

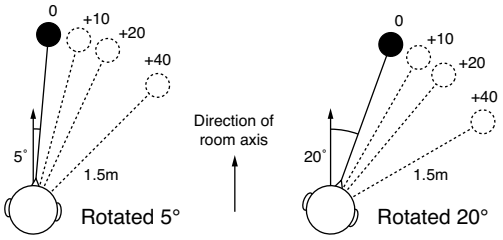

Fig. 1. Room configurations used in the experiments. The target is shown as a black circle and possible masker locations are shown as dotted circles (only three of the latter are shown).

2.2. Auditory model

The binaural input signal was processed by an auditory model [4]. Cochlear frequency analysis was approximated by a bank of 64 gammatone filters for each ear, with center frequencies spaced between 50 Hz and 8 kHz on an ERB-rate scale. Features for the recognizer were obtained by extracting the envelope from each frequency channel, which was smoothed by a first-order lowpass filter with a time constant of 8 ms, and then downsampled to a frame rate of 10 ms. The resulting spectral features were compressed by raising to the power 0.3, and then concatenated with their inter-frame differences (deltas) to provide a vector of 128 features.

For each channel f, a normalized cross-correlation $R(t,f,\tau)$ was computed between the half-wave rectified left and right-ear gammatone filter responses (denoted $x_l(t,f)$ and $x_r(t,f)$ respectively). Specifically, $R(t,f,\tau)$ was computed at 10 ms intervals of time t with a window size of 20 ms (N = 400 samples), for lags τ between -1 ms and +1 ms, as follows:

$$R(t,f,\tau)=\frac{\sum_{k=0}^{N-1} x_l(t-k,f)x_r(t-k-\tau,f)}{\sqrt{\sum_{k=0}^{N-1} x_l^2(t-k,f)}\sqrt{\sum_{k=0}^{N-1} x_r^2(t-k-\tau,f)}} \quad (1)$$

The ITD was taken to be the lag at which the largest peak occurred in the cross-correlation; this estimate was further refined by fitting a quadratic curve to the peak. The ILD was derived for each frequency channel at each time frame by computing the ratio of the energy at the output of the right- and left-ear filters, and converting to dB.

$R(t,f,\tau)$ has a range of [0,1], where 1 indicates that x_l and x_r are perfectly coherent. Following [2], we compute a measure of the interaural coherence $c(t,f)$, given by

$$c(t,f) = \max_{\tau} R(t,f,\tau). \quad (2)$$

Interaural coherence is used to identify T-F regions that are dominated by direct sound (as opposed to reflected sound), as described in Sect. 2.4.

2.3. Missing data speech recognizer

In the missing data approach to ASR, reliable and unreliable acoustic features are treated differently during decoding. The recognizer is based on hidden Markov models (HMMs) which are trained in a conventional manner using spectro-temporal features. During testing, the recognizer is provided with acoustic features and a mask; the latter indicates whether the feature describing each T-F region is reliable or not.

Clearly, the main challenge for the missing data approach is to estimate an accurate mask without prior knowledge of the target signal. Here, this is achieved by estimating masks from probability distributions of ILD and ITD. The masks estimated in this way are 'soft', i.e. each element takes a real value in the range [0,1] which indicates the probability that the element is dominated by the target [5].

Eight-state ten-mixture HMMs were used for recognition, which were trained on reverberated speech recorded at the left ear of the simulated manikin. Specifically, 4228 utterances by 55 male speakers were used, which were spatialized at 0 degrees and reverberated as described above, and then processed by the auditory model to derive training data for the recognizer. To provide a baseline for comparison, an HMM recognizer was trained with mel-frequency cepstral coefficients (MFCCs) derived from the same training set. Specifically, the feature vectors for the baseline system consisted of 12 MFCCs plus an energy term, delta and acceleration coefficients, with energy and cepstral mean normalization.

2.4. Mask estimation based on statistics of ITD and ILD

Soft missing data masks were derived from probability distributions, which indicated the probability that an observed combination of ILD and ITD was due to a source at 0 degrees azimuth. Separate distributions were trained for each frequency channel, using estimates of ILD and ITD derived from the auditory model (see Sect. 2.2). The training data consisted of 120 pairs of utterances, matched for length, for which one utterance was spatialized at 0 degrees azimuth and the other at one of eight possible azimuths (-40, -20, -10, -5, 5, 10, 20 or 40 degrees). The utterances were mixed at signal-to-noise ratios of 0, 10 and 20 dB.

Two histograms of ITD and ILD estimates were produced for each frequency channel, using bin widths of 0.01 ms and 0.1 dB respectively. The first, H_a, counted the number of observations of each ITD/ILD pair in all of the training data (i.e., observations due to the target source and masking source). The second, H_t, counted the number of observations of each ITD/ILD pair due to the target source alone. More specifically, H_t was obtained by only including observations from T-F regions that were dominated by the target, as determined from an *a priori* mask for the mixture (i.e., a binary mask constructed using prior knowledge of the target and masker signals). Separate histograms were trained for each of the 5 degree and 20 degree room configurations.

Given an observation $o = (ITD, ILD)$ from a T-F element of a target-plus-masker mixture, the probability that the

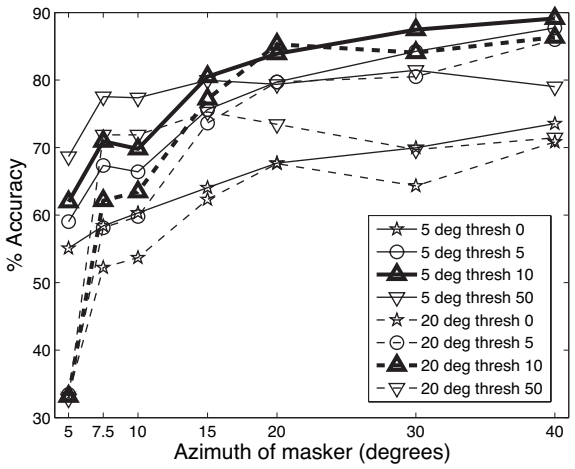

Fig. 2. Effect of histogram threshold θ_h on speech recognition accuracy. The selected threshold is shown in bold.

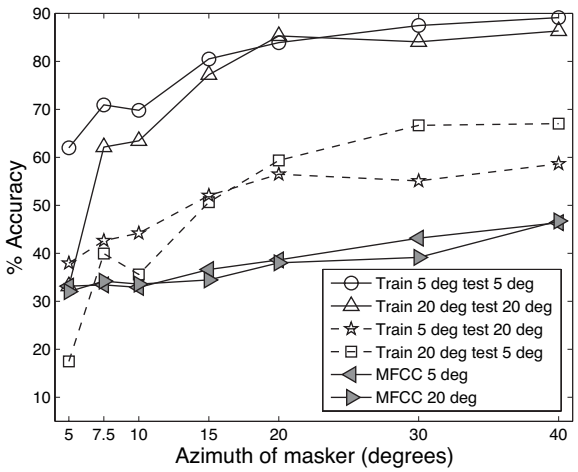

Fig. 3. Effect of room configuration on speech recognition accuracy for the missing data and baseline (MFCC) systems.

target T is dominant is given by [1]:

$$p(T|o) = p(o|T)p(T)/p(o) = H_t(o)/H_a(o) \qquad (3)$$

A soft mask was created for each test signal by identifying the ITD and ILD in each T-F region, and then applying (3) to determine the probability that the region was dominated by the target source.

A threshold was applied such that $p(T|o) = 0$ for $H_a(o) < \theta_h$, in order to reduce the effect of insufficient training data for certain combinations of ILD and ITD. The value of θ_h was derived heuristically (see Sect. 3.1). Thresholding in this way produced a cleaner estimate of $P(T|o)$ with a smooth progression between regions of low probability (in which few observations occurred) and regions of high probability (in which many observations occurred).

During training, observations from element (t, f) were discarded if the corresponding interaural coherence $c(t, f) < \theta_c$. This approach is motivated by the fact that direct sound, for which the ILD and ITD give an accurate cue to the location of the source, is associated with a high interaural coherence [2]. The threshold θ_c was derived heuristically (see Sect. 3.3). Similarly, during testing the mask value at (t, f) was set to a small number (0.3) if $c(t, f) < \theta_c$.

3. EVALUATION

The system was evaluated by measuring ASR accuracy in reverberant conditions where a speech masker was present. The test set consisted of 240 target utterances, different from those used during training of the recognizer and probability distributions, which were spatialized at 0 degrees azimuth in the 5 degree or 20 degree configuration. The masker was spatialized at azimuths of 5, 7.5, 10, 15, 20, 30 or 40 degrees, and mixed with the target source at an SNR of 0 dB. The SNR was calculated from signals spatialized at 0 degrees azimuth.

3.1. Experiment 1: Effect of histogram threshold

The histogram threshold θ_h was tuned by a series of experiments, as shown in Fig. 2. Values of $\theta_h = 0$, 5, 10 and 50 were compared for the 5 degree and 20 degree room conditions. Generally, increasing the threshold improved ASR performance. However, in some cases recognition accuracy decreased with increasing threshold, for target-masker separations of more than 20 degrees. The value of θ_h is therefore determined by the following trade-off. A low threshold allows more errors in the mask, especially when the azimuthal separation is small since ILD and ITD estimates tend to be inaccurate. A high threshold excludes some accurate ITD and ILD estimates (which tend to be associated with large azimuthal separations), but also reduces the influence of unreliable observations. In the following, we use $\theta_h = 10$.

3.2. Experiment 2: Effect of room configuration

This experiment investigated the ability of the algorithm to generalize to a room configuration that was not seen during training. Fig. 3 shows results for two conditions in which the ILD/ITD histograms were trained and tested on the same room configuration, and two conditions in which the training and testing conditions were mismatched (e.g., 'Train 20 deg test 5 deg' means that the ILD/ITD histograms were trained on the 20 degree configuration and tested on the 5 degree configuration). Clearly, the system does not generalize well when the training and testing conditions are mismatched. However, even in the mismatched condition the missing data system generally achieves a performance above the MFCC baseline.

3.3. Experiment 3: Role of interaural coherence

Figure 4 shows the effect of the interaural coherence threshold θ_c on speech recognition accuracy. A high threshold excludes

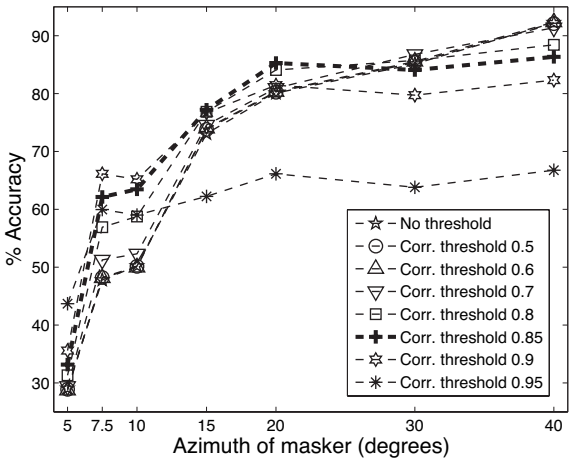

Fig. 4. Effect of interaural coherence threshold θ_c on speech recognition accuracy for the 20 degree room condition. The selected threshold is shown in bold.

most T-F regions from the mask, leading to poor ASR performance. Similarly, if T-F regions are used regardless of their interaural coherence (the 'no threshold' condition) ASR performance degrades when the azimuthal separation between the target source and the masker is large. Choosing $\theta_c = 0.85$ gave a reasonable compromise.

4. DISCUSSION

Probability distributions of ILD and ITD can be used to derive T-F masks for a missing data ASR system, thus allowing a target speaker to be recognized with good accuracy in the presence of an interfering voice and reverberation.

Our approach differs from conventional techniques for robust ASR using multiple microphones, which usually employ adaptive beamforming to derive spatially filtered acoustic features. Here, spatial information is used to *select* acoustic features rather than filter them. However, our approach is related to other sound separation techniques that exploit clustering of features in an ITD/ILD space. Roman et al. [6] describe a similar system, although it requires the number of sources, their locations and the location of the target source to be known; here, we have made the simplifying assumption that the target is at 0 degrees azimuth, and hence there is no constraint on the number of masking sources. Additionally, Roman et al. do not evaluate their algorithm in reverberant conditions, as we have done here. Yilmaz and Rickard [7] have described a blind source separation (BSS) algorithm for separating multiple sources from two acoustic inputs, which derives a binary T-F mask from the statistics of relative attenuation and inter-microphone delay. However, their goal was resynthesis of the demixed signals rather than ASR, and their algorithm assumes an anechoic mixing process which is violated in a reverberant environment.

A weakness of our approach is that performance of the algorithm is quite sensitive to the value of the histogram threshold θ_h. Using more training data, particularly for difficult conditions in which the azimuthal separation is small, would be expected to reduce the sensitivity of the system to this parameter. Additionally, we are currently investigating the use of a parametric method (Gaussian mixtures) for modeling the ILD/ITD probabilility distributions, rather than generating them directly from the training data: this will smooth the distributions and should reduce the sensitivity of the system to the training conditions.

The proposed technique is also quite sensitive to the relative placement of the source and receiver within the room. This was mitigated to some extent by labelling T-F regions as unreliable if they had a low interaural coherence, so that direct sound was given preference over reflected sound. However, there is a limit to the extent that this cue can be exploited, since the masks become very sparse if a high threshold is applied to the interaural coherence metric.

Future work will address the limitations of the algorithm discussed above. Additionally, we will investigate whether the proposed method can be used to exploit the statistics associated with other auditory features, such as those relating to periodicity pitch.

5. REFERENCES

[1] S. Harding, J. P. Barker, and G. J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Trans. Sp. Audio. Proc.*, in press 2005.

[2] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3075–3089, 2004.

[3] D. R. Campbell, *The ROOMSIM user guide (V3.3)*, 2004, http://media.paisley.ac.uk/∼campbell/Roomsim/.

[4] K. J. Palomäki, G. J. Brown, and D. L. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Sp. Comm.*, vol. 43, no. 4, pp. 361–378, 2004.

[5] J. P. Barker, L. Josifovski, M. P. Cooke, and P. D. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. ICSLP*, 2000, pp. 373–376.

[6] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *JASA*, vol. 114, no. 4, pp. 2236–2252, 2003.

[7] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Sig. Proc.*, vol. 52, no. 7, pp. 1830–1847, 2004.