A Binaural Model for Missing Data Speech Recognition in Noisy and Reverberant Conditions

Kalle J. Palomäki[1,2], Guy J. Brown[1] and DeLiang Wang[3]
[1]Department of Computer Science, University of Sheffield,
211 Portobello Street, Sheffield S1 4DP, United Kingdom.
[2]Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing
P.O. Box 3000, FIN-02015 HUT, Finland
[3]Department of Computer and Information Science and the Center for Cognitive Science,
The Ohio State University, Columbus, OH 43210, USA.
Email: kalle.palomaki@hut.fi, g.brown@dcs.shef.ac.uk, dwang@cis.ohio-state.edu

Abstract

We describe a binaural auditory model for speech recognition, which is robust in the presence of reverberation and spatially separated noise intrusions. The principle underlying the model is to identify time-frequency regions which constitute reliable evidence of the speech signal. This is achieved both by determining the spatial location of the speech source, and by applying a simple model of reverberation masking. Reliable time-frequency regions are passed to a missing data speech recogniser. We show, firstly, that the auditory model improves recognition performance in various reverberation conditions when no noise intrusion is present. Secondly, we demonstrate that the model improves performance when the speech signal is contaminated by noise, both for an anechoic environment and in the presence of simulated room reverberation.

1. Introduction

Human listeners are able to recognise speech even in noisy acoustic environments. This remarkable robustness is due to two main factors. Firstly, mechanisms of speech perception are largely unaffected when the speech signal is distorted or masked by other sounds. Secondly, human listeners are able to perceptually segregate a target sound from an acoustic mixture. In contrast, automatic speech recognition (ASR) in noisy acoustic environments remains very problematic. It is reasonable to argue, therefore, that ASR performance could be improved by adopting an approach that models auditory processing more closely. Additionally, such auditory models may contribute to our understanding of human hearing by clarifying the computational processes involved in speech perception.

The term *auditory scene analysis* (ASA) has been introduced to describe the process by which listeners parse an acoustic mixture [5]. In this process, acoustic components that are likely to have arisen from the same environmental event are grouped to form a perceptual stream. Streams are subjected to higher-level processing, such as language understanding. Auditory grouping is known to exploit physical characteristics which are related to common spectro-temporal properties of sound. Additionally, ASA uses information about the spatial location of sound sources, which is principally encoded by interaural time difference (ITD) and interaural intensity difference (IID) cues at the two ears. Indeed, the role of spatial location in sound separation has been appreciated since the early fifties [16].

The problem of segregating speech from a noisy background has been investigated over the last decade using computational approaches to ASA (see [15] for a review). However, in much of this work binaural hearing has been neglected in favour of simpler monaural mechanisms (notable exceptions include [4], [11] and [13]). It is also apparent that few computational approaches to ASA have been evaluated in reverberant conditions, presumably because of the difficulty of the task.

Recently, progress has also been made in developing ASR systems that exploit principles of human speech perception. Cooke and his co-workers [9] have interpreted the robustness of speech perception mechanisms in terms of their ability to deal with 'missing data', and have proposed an approach to ASR in which a hidden Markov model (HMM) classifier is adapted to deal with missing or unreliable features. The missing data paradigm is complementary to computational ASA; an auditory model can be used to decide which acoustic components belong to a target speech source, and only these 'reliable' features are passed to the recogniser. In this study, we propose a binaural approach to computational ASA, and show that it provides an effective front-end for missing data recognition of speech in noisy and reverberant environments. The present study extends previous work in this field (e.g., [11], [13]) which has described a binaural front-end for speech recognition, but has not evaluated it in the presence of reverberation.

2. Model

The model (Figure 1) is divided into monaural and binaural pathways. The monaural pathway is responsible for peripheral auditory processing, and for producing feature vectors for the speech recogniser. It also implements a reverberation masking model, which improves recognition performance in reverberant conditions. The binaural pathway is responsible for sound localisation and separation according to common azimuth.

2.1. Monaural pathway

In the first stage of the monaural pathway, the direction dependent filtering effects of the pinna, head and torso are modelled by convolving the acoustic input with a head-related impulse response (HRIR) for each ear. The set of HRIRs used in this study were measured from the KEMAR artificial head [10]. Cochlear frequency analysis is simulated by a bank of 32 bandpass gammatone filters with centre frequencies spaced on the equivalent rectangular bandwidth (ERB) scale. The output of each filter is half-wave rectified and compressed to give a representation of auditory nerve activity.

A second monaural processing pathway is needed to provide feature vectors for the speech recogniser. First, the

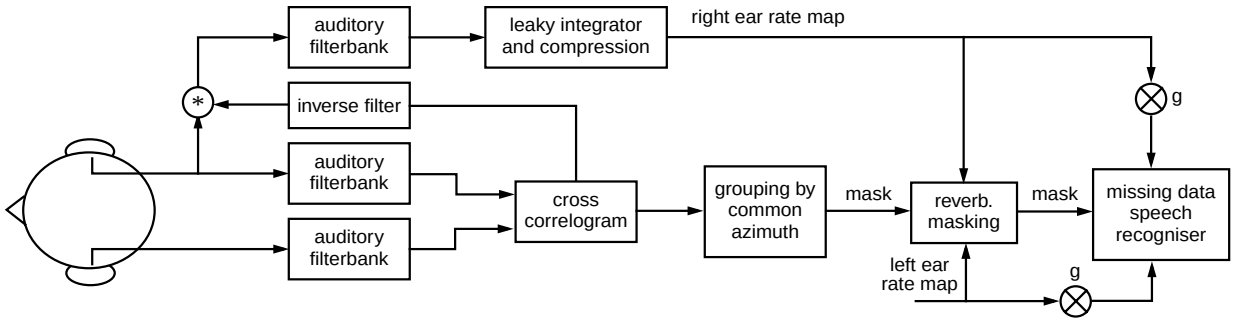

Figure 1: Schematic diagram of the right ear of the model.

convolutional distortion caused by the HRIR filtering must be compensated, otherwise it will degrade the performance of the HMM recogniser. Compensation is performed by an all-pole inverse filter [14], which is derived from a phase minimized version of the HRIR corresponding to the model's estimate of the azimuth of the speech source. Subsequently, the inverse filtered signal is fed through another gammatone filterbank and the instantaneous Hilbert envelope is computed at the output of each filter [8]. This is smoothed by a first-order low-pass filter with an 8 ms time constant, sampled at 10 ms intervals, and finally cube root compressed to give an auditory firing rate representation. Rate maps computed for the left and the right ears are averaged. Because reverberation introduces level changes which degrade recogniser performance, a gain adjustment g was applied to the rate maps. We use g=1 for the non-reverberant case, g=0.9 for reflection factors between 0.5 and 0.7, and g=0.86 for reflection factors between 0.8 and 0.95.

2.2. Binaural pathway

In the binaural pathway, the model derives an estimate of ITD by computing a cross-correlogram. Given the left and right ear auditory nerve activity in channel i at time step j, $l(i,j)$ and $r(i,j)$, the cross correlation for delay τ is

$$C(i, j, \tau) = \sum_{k=0}^{M-1} l(i, j-k)r(i, j-k-\tau)w(k) \qquad (1)$$

where w is a rectangular window of width M time steps. We use M=600, corresponding to a window duration of 30 ms, and consider values of τ between ±1 ms. Computing $C(i,j,\tau)$ for each channel i gives a cross-correlogram, which is computed at 10 ms intervals. Each cross-correlation function is then mapped from ITD to an azimuth scale using a lookup table, giving a function $C(i,j,\phi)$, where ϕ is azimuth in degrees.

Subsequently, a 'skeleton' is formed for each function. In this technique, each peak in the cross-correlation function is replaced with a gaussian whose width is narrower than the original peak, and is proportional to the channel centre frequency. This process is similar in principle to lateral inhibition, and leads to a sharpening of the cross-correlogram. The skeleton cross-correlation functions $S(i,j,\phi)$ are summed over frequency to give a pooled cross-correlogram, in which the location of each sound source is indicated by a clear peak.

2.3. Missing data speech recogniser

The speech recognition stage of our system employs the missing data technique [9]. In general, the classification problem in speech recognition involves the assignment of an acoustic observation vector v to a class C. However, if a noise intrusion is present some components of v may be unreliable or missing. In such cases, the likelihood $f(v|C)$ cannot be computed in the normal manner. The 'missing data' technique addresses this problem by partitioning v into reliable and unreliable components, v_r and v_u. The reliable components v_r are directly available to the classifier. In the simplest approach, the components of the unreliable part v_u are simply ignored so that classification is based on the marginal distribution $f(v_r|C)$. However, when v is an acoustic vector additional constraints can be exploited, since it is known that uncertain components will have bounded values (the 'bounded marginalisation' method [9]). Here, we use bounded marginalisation in which v is an estimate of auditory nerve firing rate, so the lower bound for v_u is zero and the upper bound is the observed firing rate.

In practice, a binary 'mask' $m(i,j)$ is used to indicate whether the acoustic evidence in each time-frequency region is reliable. Here, mask values are determined by two heuristics; common azimuth and reverberation masking.

2.4. Grouping by common azimuth

The first heuristic implements auditory grouping by common azimuth. The azimuths of the speech and noise, ϕ_s and ϕ_n, are derived from the pooled skeleton cross-correlogram. We assume that $\phi_s > \phi_n$ (i.e., that the speech lies to the right of the noise). Values in the mask are then set according to

$$m(i, j) = \begin{cases} 1 \text{ if } C(i, j, \phi_s) > C(i, j, \phi_n) \\ 0 \qquad \text{otherwise} \end{cases} \qquad (2)$$

When the speech and noise are spatially separated, this heuristic effectively identifies the time-frequency regions that are dominated by the speech source.

Some of the experiments reported in Section 3 consider the recognition of reverberated speech when no noise intrusion is present. In such conditions, ϕ_n corresponds to the azimuth of the strongest reflection of the speech source. Hence, although it was conceived as a mechanism for auditory grouping rather than dereverberation, (2) will reduce the effect of reverberation by emphasising acoustic components that originate directly from the azimuth of the speech source.

2.5. Reverberation masking

A second heuristic is applied to the averaged rate map which implements a simple mechanism for forward masking. This emphasizes direct sound and attenuates parts of the spectrum that are contaminated with reverberation. First, the rate map

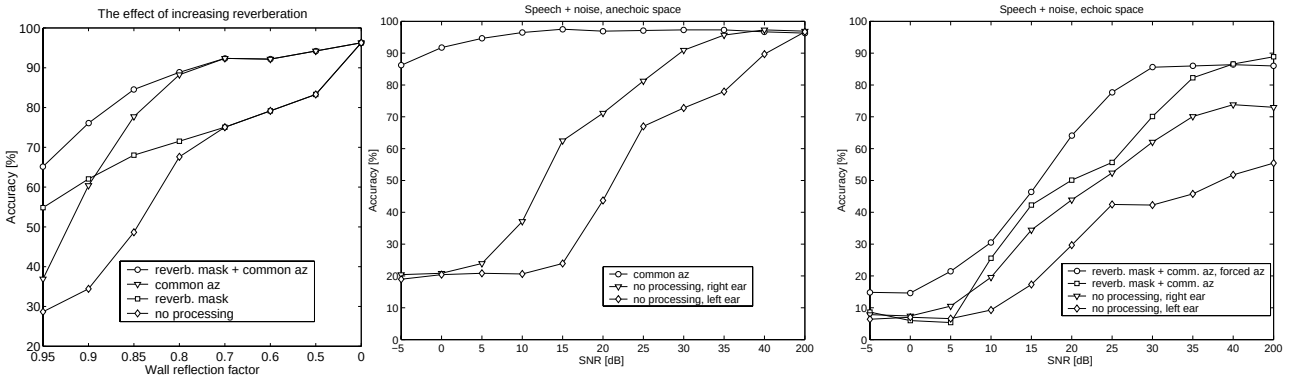

Figure 2. Left: Speech recognition score as a function of reverberation. Middle & Right: Speech recognition performance in the presence of a noise intrusion in anechoic conditions (middle) and in a simulated room (right) with a wall reflection factor of 0.8.

value $a(i,t_0)$ at every time instant t_0 in channel i is considered as a masker. Starting from this value, a monotonically decreasing masked threshold function $\theta(i,t)$ is computed:

$$\theta(i,t) = \lambda a(i,t_0) - \kappa \cdot (t - t_0) \quad (3)$$

where t is current time instant, forward in time compared to t_0. The values of the mask are then decided according to:

$$m(i,j) = \begin{cases} 1 \text{ if } a(i,t) > \theta(i,t) \\ 0 \quad \text{otherwise} \end{cases} \quad (4)$$

The values of the parameters λ and κ were the same for the all frequency channels in a single experimental case. These parameters were experimentally tuned to each reverberation condition. The optimal masked threshold function depended on the length of the room impulse response, so that with increasing length more speech samples were judged unreliable (i.e. $m(i,j)=0$). It is important to note that this principle does not correspond exactly to the temporal masking phenomenon which occurs in the auditory system. The model proposed here is intended to remove parts of the signal which are contaminated by reverberation, but such regions might still be audible to listeners (i.e., not masked). Auditory post-masking related to room reverberation is further discussed in [7].

3. Evaluation

We evaluated the model using a variety of noise conditions. In the first experimental case, the effect of reverberation on recognition performance was investigated by using a mirror image model of room acoustics during stimulus generation [1]. In the second experiment we evaluated speech recognition performance in the presence of a spatially separated intrusive noise, under anechoic and reverberant conditions.

3.1. Corpus & noise

The model was evaluated on a subset of male speakers from the TiDigits connected digits corpus [12]. Auditory rate maps were obtained for the training section of the corpus, and were used to train 12 word-level HMMs (a silence model, 'oh', 'zero' and '1' to '9'). All models were trained on unreverberated signals. A subset of 100 male utterances from the test set of the corpus were used for evaluating the model.

The rock music noise intrusion from Cooke's corpus [8] was used to test the model. The amplitude of the noise signal was scaled to give a range of signal-to-noise ratios (SNRs) from -5 dB to 40 dB. The noise intrusion and test utterance were then convolved with left ear and right ear HRIRs corresponding to angles of incidence of -30 degrees for the noise and 10 degrees for the speech. In experiments where reverberation was present, the HRIRs incorporated a room impulse response which was generated using the image model (see below). The spatialised noise and utterance signals were then summed for each ear, giving a binaural mixture.

3.2. Mirror image reverberation model

The basic principle of the image model is that reflection paths from a sound source to a listener are found by reflecting the sound source against all surfaces of the room [1]. Here, we created a small rectangular room (length x=6m, width y=4m and height z=3m) to mimic a small office. The amount of reverberation was adjusted by varying the wall reflection parameter. We positioned the listener in the middle of the floor (x=3m, y=2m, z=2m) and presented speech and noise from a distance of 1.5 m at different horizontal angles. To model the interaction of sound waves with the listener's head and torso, each reflection was convolved with a HRIR corresponding to the direction of the reflection. The delay and sound attenuation were set according to the distance between the image source and the listener. For simplicity, the azimuth and elevation of each reflection were quantised to fit the resolution of the KEMAR HRIR data (see [10]). The elevation angle was rounded to the nearest 10 degrees in the interval -40 to 90 degrees, and larger negative values were always rounded to -40. The azimuth resolution was 5 degrees in the vicinity of the horizontal plane and decreased as the elevation increased to higher positive or negative angles.

4. Results

Firstly, the effect of reverberation was investigated in the absence of a noise intrusion. The left panel of Figure 2 shows the recognition performance using missing data with different mask estimation heuristics; monaural reverberation masking alone (□), grouping according the common azimuth (▽) and grouping by common azimuth combined with reverberation masking (○). The figure also shows recognition performance without missing data processing (◇). All the result graphs were obtained using the averaged left and right ear rate maps.

Grouping by common azimuth gives a substantial improvement in recognition accuracy when combined with the missing data recogniser. Addition of the reverberation masking model further improves recognition performance in conditions where the wall reflection factor exceeds 0.8.

The middle and right panels of Figure 2 compare recognition performance in anechoic and reverberant conditions, when a spatially separated noise intrusion is present at varying SNRs. Importantly, the missing data approach based on the averaged left and right ear rate maps (○) performs significantly better than a strategy that only uses the ear nearest to the sound source (▽). As expected, recognition performance is poorest using the ear nearest to the noise (◇).

In the reverberant case (right panel of Figure 2) the performance of the model was poor at low SNRs (□), although it still generally exceeds that of the recogniser without preprocessing (◇ and ▽). To identify the cause of this performance drop, we considered a condition in which the azimuth estimate was forced to the proper value (10 degrees). When this was done, grouping according to common azimuth worked well (○), indicating that our mechanism for identifying the azimuths of the speech and noise sources becomes unreliable at low SNRs in the presence of reverberation.

In summary, in the anechoic case our system gives a substantial increase (up to 70%) in recognition rate over a -5 to 30 dB SNR range. In the reverberant case the auditory front-end appears to improve recognition performance most clearly at higher SNRs (25-200 dB).

5. Discussion

Overall, using binaural processing to estimate masks for missing data recognition provides some clear benefits over a monaural approach. For example, monaural algorithms that segregate concurrent sounds according to their fundamental frequencies fail in unvoiced regions of speech [6].

The results in Figure 2 demonstrate the capabilities of our model to deal with reverberation. Using our approach around 90% recognition rate was achieved in realistic office room conditions (0.7, 0.8). However, when the reflection ratio was increased the results dropped rather rapidly. Reflection factors of 0.9 and 0.95 correspond to around 0.5 sec T60 reverberation time, which is widely agreed to be good value for designing lecture halls. In these cases our results already dropped down to 70-80% (however, they represent a 40% increase in performance compared to recognition without missing data processing). These facts clearly demonstrate the sensitivity of HMM-based speech recognition to room reverberation. In this study HMM models were trained in anechoic conditions. It would be interesting to compare the difference between the missing data approach presented here and recognition from models trained in reverberant conditions.

In the experiment where a noise intrusion was mixed with speech in reverberant conditions, the sound localization model failed to produce accurate azimuth estimates. This is not surprising, since the precedence effect was not considered in this study. Even a simple implementation of the precedence effect is likely to improve azimuth estimation; we will include such processing in future versions of the model.

Barker et al. [3] showed that instead of using binary masks, better results can be obtained by making soft decisions where the mask values can be real numbers between 0 and 1. Missing data using soft decisions is likely to further improve the performance of the model, and our model of reverberation masking could easily be adapted to work with real-valued masks. Additionally, a multi-source decoding principle [2] could provide further improvements in performance by indicating whether the activity at the particular region of auditory space is speech-like.

Finally, the experiments described here simulated an indoor acoustic environment in which speech and rock music were presented from spatially separated loudspeakers. Of course, this is not particularly representative of real-world listening conditions; typically, environmental noise does not originate from a single location, but is distributed in auditory space. Future work will consider more challenging acoustic environments of this kind.

Acknowledgement

The project was funded by the EC TMR SPHEAR project and partially supported by Finnish Tekniikan edistämissäätiö grant.

References

[1] J. B. Allen and D. A. Berkley (1979) Image method for efficiently simulating small-room acoustics, *J. Acoust. Soc. Am.*, 65, pp. 943-950.

[2] J. Barker, M. P. Cooke and D.P.W. Ellis (2000) Decoding speech in the presence of other sound sources, *Proc. ICSLP'00*, IV, pp. 270-273.

[3] J. Barker, M. P. Cooke, L. Josifovski and P. D. Green (2000) Soft decisions in missing data techniques for robust automatic speech recognition, *Proc. ICSLP'00*, I, pp. 373-376.

[4] M. Bodden (1993) Modeling human sound-source localization and the cocktail-party-effect. *Acta Acustica*, 1, pp. 43-55.

[5] A. S. Bregman (1990) *Auditory scene analysis*. MIT Press.

[6] G. J. Brown, D. L. Wang and J. Barker (2001) A neural oscillator sound separator for missing data speech recognition. *Proc. IJCNN-01*, in press.

[7] J. M. Buchholz, J. Mourjopoulos, J. Blauert (2001) Room Masking: Understanding and Modelling the Masking of Room Reflections, *110th AES convention*, preprint Nr. 5312.

[8] M. P. Cooke (1993) *Modelling auditory processing and organization.* Cambridge, UK: Cambridge University Press.

[9] M. P. Cooke, P. D. Green, L. Josifovski and A. Vizinho (2001) Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data, *Speech Communication*, 34, pp. 267-285.

[10] B. Gardner and K. Martin (1994) HRTF measurements of KEMAR dummy-head microphone, Technical Report #280, MIT Media Lab URL: ftp://sound.media.mit.edu/pub/Data/KEMAR/.

[11] H. Glotin, F. Berthommier and E. Tessier (1999) A CASA-labelling model using the localisation cue for robust cocktail-party speech recognition, *Proc. EUROSPEECH'99*, pp. 2351-2354.

[12] R. G. Leonard, (1984) A database for speaker-independent digit recognition, *Proc. ICASSP'84*, pp. 111-114.

[13] H. G. Okuno, T. Nakatani and T. Kawabata (1999) Listening two simultaneous speeches, *Speech communication, 27,* pp. 299-310.

[14] A. V. Oppenheim and R. W. Schafer (1999) *Discrete-time signal processing* (2nd edition). Englewood Cliffs, NJ:Prentice Hall.

[15] D. F. Rosenthal and H. G. Okuno (1998) *Computational auditory scene analysis*, Lawrence Erlbaum Associates Publishers, New Jersey.

[16] W. Spieth, W. Curtis and J. C. Webster (1954) Responding to one of two simultaneous messages, *J. Acoust. Soc. Am.* 26, pp. 391-396.