A computational model of the speech reception threshold for laterally separated speech and noise

Guy J. Brown[1] *and Kalle J. Palomäki*[1,2,3]

[1]Department of Computer Science, University of Sheffield, United Kingdom
[2]Lab. of Acoustics and Audio Signal Processing, Helsinki Univ. of Technology, Finland
[3]Apperception & Cortical Dynamics (ACD), Dept. of Psychology, University of Helsinki, Finland
g.brown@dcs.shef.ac.uk, kalle.palomaki@hut.fi

Abstract

Recent psychophysical studies suggest that human listeners do not segregate concurrent sounds by grouping frequency regions that have a common interaural time difference (ITD). However, such an approach is adopted by most computational auditory scene analysis (CASA) systems that use binaural cues. Here, we propose a CASA system that separates a target speech signal from a noise interferer, but does not require the ITD of the two sources to be consistent across frequency. We compare the CASA system with human performance on the same task, in which the speech reception threshold (SRT) is measured for speech and noise stimuli which have consistent or inconsistent ITDs in different frequency bands. The CASA system is shown to be in qualitative agreement with human performance.

1. Introduction

It is well known that listeners are better able to recognise speech in the presence of a noise masker if the speech and noise originate from different locations in space. This observation has motivated a number of computational auditory scene analysis (CASA) systems, which use binaural cues to segregate a target speech signal from spatially separated noise [1], [2], [3].

An assumption made by these computational systems is that listeners segregate sound sources by grouping frequency regions which share a common interaural time difference (ITD). However, psychophysical studies suggest that human listeners do not adopt this strategy when segregating concurrent sounds. A recent illustration of this is provided by Edmonds [4]. In his experiment, target speech and an interfering sound were split into high and low frequency bands and presented in three ITD configurations ('same', 'consistent' and 'swapped'), as shown in Fig. 1. Using a speech reception threshold (SRT) test, Edmonds confirmed that speech intellibility was improved in the 'consistent' condition compared to the 'same' condition. However, he found no difference in intelligibility between the 'consistent' and 'swapped' conditions. This result is incompatible with a mechanism based on grouping by common ITD, which should fail badly in the 'swapped' condition due to inappropriate grouping of speech and interferer bands that share the same ITD. It therefore appears that listeners can exploit a difference in ITD between speech and noise, but it is not necessary for this difference to be consistent across frequency.

An accurate model of human performance should be able to replicate Edmonds' findings, but as already noted the majority of binaural CASA systems do not. Here, we address this deficiency by proposing a CASA system which exploits the difference in ITD between a target speech source and noise interferer in independent frequency bands. The system is based on the 'missing data' framework for automatic speech recognition (ASR), and we evaluate the system by using it as a 'subject' in Edmonds' SRT test.

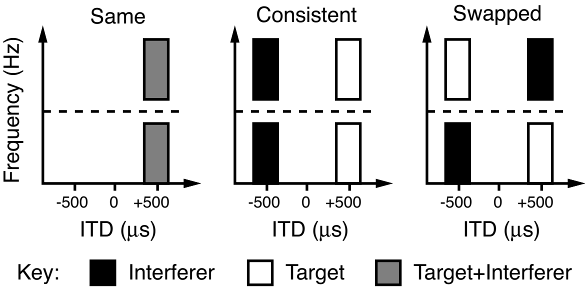

Figure 1: Schematic of Edmonds' swapped ITD experiment (adapted from Figure 3.8 in [4]). The target and interferer are split into two frequency bands and presented with the same ITD, consistent ITD at opposite sides, or swapped ITD at opposite sides. The splitting frequency is denoted by a dotted line.

2. Perceptual experiment

In order to compare our system directly with psychophysical findings, we first replicate Edmonds' SRT experiment using speech data which is more suited to our ASR system (i.e., we use spoken digits rather than the Harvard Sentence List used by Edmonds [4]).

2.1. Corpus

The utterances employed in the SRT test were selected from the test set of the TiDigits connected digits corpus [5] according to a number of criteria, which aimed to ensure that all trials would be of equal difficulty. Firstly, the number of syllables in each utterance was balanced by selecting only those four digit utterances for which each digit contained a single syllable. Hence the digits 'oh', 'one' to 'six', 'eight' and 'nine' were used, and 'zero' and 'seven' were omitted. The suitability of monosyllabic digits for estimating the SRT has previously been verified by Ramkissoon *et al.* [6]. All utterances were verified for approximately equal intelligibility by informal listening tests conducted by the authors. Talkers were drawn from a limited number of accent groups, and those with atypical accents or whose speech exhibited large variations in pitch or intensity were excluded. The sampling rate of the speech signals was 20 kHz.

Speech-shaped noise was generated by passing Gaussian noise through a FIR filter. The filter was designed to have the same magnitude response as the long-term spectrum of the selected TiDigits utterances.

2.2. Method

Twelve native English speaking subjects, all of whom reported normal hearing, participated in a replication of Edmond's SRT experiment using spoken digits. From the speech material described above, 6 lists were constructed for the SRT test, each of which contained 19 utterances. Within each list, the utterances originated from 19 different speakers and the order of speakers was held constant across the lists. Four additional lists of 9 utterances were constructed for familiarising the subjects with the test procedure. All utterances were presented only once to each subject to prevent memorisation of the speech material.

Experiments were conducted with a subset of the stimuli used by Edmonds [4]. The target speech signal and speech-shaped noise were split into two frequency bands, with splitting frequencies at 750 Hz or 1500 Hz (see Fig. 1). The two bands were separated by a gap of one ERB, centered on the splitting frequency.

Three configurations of the speech and noise were used. In the 'same' configuration, the speech and noise were presented with an ITD of +500 μs (i.e., to the right side of the head). In the 'consistent' configuration, the speech and noise were presented on different sides of the head, with ITDs of +500 μs and −500 μs respectively. Finally, in the 'swapped' configuration the low frequency band of the speech and the high frequency band of the noise were presented with an ITD of +500 μs, and the low frequency band of the noise and the high frequency band of the speech were presented with an ITD of −500 μs.

The three ITD configurations and two splitting frequencies gave a total of six experimental conditions. Since the noise masker was not identical in all conditions, it was not possible to balance the difficulty of different utterance lists by adjusting the initial noise level. Instead, the sequence of experimental conditions (which was initially chosen randomly) was rotated for each listener [4]. To achieve this, the 12 subjects were divided into two groups of six, in order to match them with six experimental conditions and six utterance lists. Each subject heard the utterance lists in the same order, but the experimental conditions were rotated so that subject 1 in each group heard condition 1 first, subject 2 in each group heard condition 2 first, and so on.

In all tests, the noise was presented at a constant level of 70 dB SPL. Prior to the SRT test for each experimental condition, the speech was initially presented at a level at which it was completely masked by the noise (the corresponding SNR was −26 dB). The level of the speech was then incremented in steps of 4 dB until the subject achieved 50% recognition accuracy. This procedure was repeated for two different utterances, and the average SNR obtained on the two attempts was used as a starting point for the SRT test.

In the SRT test itself, the level of the speech was adjusted adaptively using a 1 up / 1 down tracking procedure [7]. If subjects achieved a recognition accuracy of 75% then the level of the next utterance was reduced by 2 dB, otherwise the level of the speech was increased by 2 dB. The SRT for each subject and experimental condition was achieved by averaging the SNRs obtained after each level adjustment, with the exclusion of the one originating from the initial level calibration.

The experiments were performed in an IAC single-walled soundproof room. Stimuli were presented to subjects via a Tucker-Davis RP 2.1 headphone driver and Sennheiser HD 580 headphones. Subject's responses were collected via a computerised test procedure. No corrections or replications were allowed once the subject had answered by typing on a computer keyboard. Before the actual SRT tests, subjects practised the test procedure over four lists of nine utterances, in which the speech and noise were presented diotically without ERB gaps.

2.3. Results

Data from the psychoacoustic test were analysed using repeated measures ANOVA with a two-way design. Effects of ITD condition (same, consistent or swapped) and splitting frequency (750 Hz or 1500 Hz) were investigated. The ITD condition had a statistically significant effect on the SRT ($F[2, 22] = 359.91; P < 0.001$), whereas the effects of splitting frequency and the interaction between ITD condition and splitting frequency were nonsignificant (P = n.s.). Tukey HSD Post hoc analyses for the ITD condition revealed statistically significant differences in all comparisons ($P < 0.001$), where the SRTs were −16.6 dB (SEM 0.29 dB) for consistent ITD, −15.4 dB (SEM 0.16 dB) for swapped ITD, and −9.9 dB (SEM 0.20) for same ITD.

Edmonds [4] did not find a significant difference between the swapped and consistent ITD conditions, as we do here. Since our test procedures were essentially the same, it is possible that our experiment was more discriminative because of differences in the subjects and speech material that were used. However, we confirm the key finding from Edmonds' experiment: listeners gain a substantial benefit from a difference in ITD between the speech and the noise, regardless of whether the ITD is consistent over frequency or not.

3. Computational model

The computational model consists of three stages; peripheral frequency analysis, selection of acoustic features using binaural and fundamental frequency (F0) cues, and speech recognition by a 'missing data' ASR system.

3.1. Missing data recognition

The automatic speech recognition component of the model utilises the 'missing data' technique with bounded marginalisation [8]. In this approach, acoustic features are treated differently during decoding depending on whether they are labelled as reliable or unreliable evidence for the target speech source. In practice, the recogniser is supplied with acoustic features and a binary mask, in which values of zero and one indicate unreliable and reliable features respectively. Here, acoustic features are provided by a model of the auditory periphery, and the mask is determined by binaural cancellation and monaural F0-based grouping.

Acoustic features were computed for the training section of the TiDigits corpus [5], and used to train a silence model and eleven word-level hidden Markov models (HMMs) as in our previous study [3]. The models for 'zero' and 'seven' were not used during testing. Each HMM consisted of 8 no-skip, straight-through states with observations modelled by a 10-component diagonal covariance Gaussian mixture. All models were trained on clean speech.

3.2. Peripheral model

Peripheral auditory processing is modelled by two banks of bandpass ‘gammatone’ filters. For each ear, $N = 32$ filters are used, with centre frequencies uniformly spaced between 50 Hz and 8 kHz on an ERB-rate scale. To provide acoustic features for the recogniser, the instantaneous Hilbert envelope is computed at the output of each filter, and smoothed by a first-order lowpass filter with a time constant of 8 ms. The smoothed envelope is then sampled at 10 ms intervals and compressed by raising it to the power of 0.3.

For subsequent computation of F0 and binaural cues, a crude simulation of auditory nerve (AN) activity is also obtained from each filter by half-wave rectifying its output. Here, we denote the AN activity for the left and right ears as $a_L(t, f)$ and $a_R(t, f)$ respectively, where t indexes time and f is the frequency channel.

3.3. Mask estimation

The mask for missing data ASR is estimated using a combination of two approaches. Firstly, a difference in ITD between the speech and the noise is exploited by a binaural EC mechanism, which estimates an ‘EC mask’ based on the amount of speech energy present after cancellation of the noise background. Secondly, F0-based grouping is used to derive a ‘pitch mask’, in which harmonics of the target speech signal are selected. Note that in the consistent and swapped conditions of Edmonds’ experiment, the model is able to exploit both ITD and F0 cues. In the ‘same’ condition, no difference in ITD is present and hence the model relies on F0-based grouping only.

The F0 of the speech signal is estimated by computing the autocorrelation of the simulated auditory nerve response for each frequency channel. The resulting ‘correlogram’ for each ear $e \in \{L, R\}$ is given by

$$acf_e(n, f, \tau_1) = \sum_{t=0}^{T-1} a_e(n-t, f)a_e(n-t-\tau_1, f)w(t) \qquad (1)$$

where n indexes the time frame, τ_1 is the autocorrelation lag and $w(t)$ is a rectangular window of width 20 ms (i.e., $T = 400$ samples). A composite correlogram is then obtained by summing over both ears,

$$acf(n, f, \tau_1) = \sum_{e\in\{L,R\}} acf_e(n, f, \tau_1) \qquad (2)$$

and a ‘summary’ correlogram is computed by pooling over all frequency channels:

$$s(n, \tau_1) = \sum_{f=1}^{N} acf(n, f, \tau_1) \qquad (3)$$

The pitch period $p(n)$ corresponds to the lag τ_1 at which the highest peak occurs in the summary correlogram,

$$p(n) = \underset{\tau_1}{\operatorname{argmax}} \; s(n, \tau_1) \qquad (4)$$

For each frame n, we determine whether a voice pitch is present by applying a threshold θ_v to $s(n, p(n))$. The latter is normalised by the energy in the frame (i.e., the value of the summary correlogram at zero lag). If the frame is unvoiced, then all elements in the pitch mask $m_p(n, f)$ for that frame are set to zero. Otherwise, the mask element for channel f is set to one if the activity in the corresponding correlogram channel exceeds a threshold value θ_p at the lag $p(n)$:

$$m_p(n, f) = \begin{cases} 1 & \text{if } \hat{s}(n, p(n)) > \theta_v \wedge acf(n, f, p(n)) > \theta_p \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

$$\hat{s}(n, p(n)) = s(n, p(n))/s(n, 0) \qquad (6)$$

On the basis of experiments with a small validation set, we set $\theta_v = 0.75$ and $\theta_p = 0.75$.

Binaural signal detection is performed by equalisation-cancellation [9]. In the equalisation stage, the left and right binaural signals in each frequency band are equalised by dividing by their r.m.s. values, computed over the analysis window $w(t)$. Cancellation is then performed on the equalised inputs $\hat{a}_L(t)$ and $\hat{a}_R(t)$ by

$$ecf(n, f, \tau_2) = \sum_{t=0}^{T-1} |\hat{a}_L(n-t, f) - \hat{a}_R(n-t-\tau_2, f)|w(t) \qquad (7)$$

where τ_2 is the interaural delay. By analogy with the correlogam described above, we refer to $ecf(n, f, \tau_2)$ as a ‘cancellogram’. The ITD of the noise background is estimated by pooling the cancellogram over the first ten frames of the acoustic signal, in which it is known that there is no speech energy:

$$ecf_{noise}(f, \tau_2) = \sum_{n=1}^{10} ecf(n, f, \tau_2) \qquad (8)$$

The ITD of the noise background is then determined separately for each frequency channel as follows:

$$ITD_{noise}(f) = \underset{\tau_2}{\operatorname{argmin}} \; ecf_{noise}(f, \tau_2) \qquad (9)$$

The EC mask is then set to unity in those time-frequency regions where the residue from cancellation at the delay of the noise exceeds a threshold value, i.e.

$$m_{ec}(n, f) = \begin{cases} 1 & \text{if } ecf(n, f, ITD_{noise}(f)) > \theta_{ec} \\ 0 & \text{otherwise} \end{cases} \qquad (10)$$

The parameter θ_{ec} was set to 25.0 (model units) by inspection. Finally, the mask for the recogniser $m(n, f)$ is obtained by combining the pitch mask $m_p(n, f)$ and EC mask $m_{ec}(n, f)$ using logical OR.

3.4. Across-channel processing

The above model employs a ‘within channel’ approach to noise cancellation, in that the ITD of the noise is estimated for each frequency channel independently. For comparison, we also consider an ‘across channel’ approach which exploits the coherence of ITD across frequency in order to segregate the target speech from the background noise. As such, it is typical of a general class of across-channel models including those of [2], [3].

The across-channel approach works as described above, except that the interaural delay of the noise, ITD_{noise}, is estimated from a summary cancellogram given by

$$ecf_{noise}(\tau_2) = \sum_{n=1}^{10} \sum_{f=1}^{N} ecf(n, f, \tau_2) \qquad (11)$$

The number of minima that lie below the mean of $ecf_{noise}(\tau_2)$ is determined. If one minimum is found (which would be expected in the ‘same’ condition of Edmonds’ experiment), then ITD_{noise} is set to the corresponding value of τ_2. If two minima are located, then ITD_{noise} may be set to either of the corresponding interaural delays depending on the experimental condition. In the ‘consistent’ condition, the noise is associated with a negative value of τ_2 (see Fig. 1). In the ‘swapped’ condition, the high-frequency and low-frequency parts of the noise are associated with positive and negative values of τ_2 respectively. Mask estimation then proceeds as above, except that the criterion for setting a mask element to unity in (10) is $ecf(n, f, ITD_{noise}) > \theta_{ec}$, i.e. it is assumed that the noise has

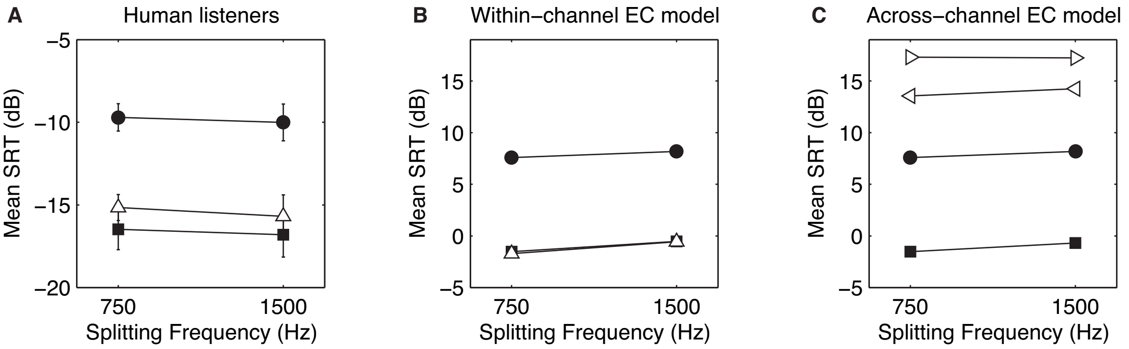

Figure 2: **A**. Human performance on the SRT test for the ‘same’ (●), ‘consistent’ (■) and ‘swapped’ (△) ITD conditions. Error bars correspond to +/- one standard deviation. **B**. Performance of the within-channel EC model on the SRT test. Conditions are labelled as in the previous figure. **C**. Performance of the across-channel EC model. Conditions are same (●), consistent (■), and swapped ITD with selection of the negative lag (▷) or positive lag (◁) in the summary cancellogram.

the same ITD in all frequency channels f.

3.5. Results

The MATLAB scripts used to run the SRT test described in Sect. 2.2 were slightly modified to enable the computational model to act as a ‘subject’ in the test. Specifically, acoustic features and a mask were derived for the speech signal presented on each trial, and these were decoded by the missing data recogniser. The resulting transcription was scored in the same way as listeners responses, and the SNR was adjusted adaptively as before.

Results are shown in Fig. 2. Overall, the digit recognition performance of the computational model is poorer than that of human listeners, as indicated by its substantially higher SRT. For example, in the baseline condition the model SRT is about 15 dB greater than the human SRT. However, in terms of relative performance in the three experimental conditions, the within-channel EC model (panel B) provides a good match to human data. Specifically, the within-channel EC model gains the same SRT advantage from a difference in ITD between the speech and noise, regardless of whether the ITDs are consistent across frequency or swapped. The performance of the across-channel EC model is shown in panel C of the figure. In the swapped condition, the SRT was lower when the rightmost dip in the summary cancellogram was selected rather than the leftmost dip (i.e., speech recognition performance was better when the low-frequency part of the speech was selected as reliable in the mask; see Fig. 1). However, the performance of the across-channel EC model is seriously disrupted by swapping the ITDs between frequency bands, and therefore does not correctly model human performance.

4. General discussion

We have shown that a binaural CASA system which exploits within-channel differences in ITD between two sound sources can qualitatively model psychophysical data relating to the perception of spatially separated speech and noise mixtures. An interesting feature of the proposed model is that it employs both cancellation (by ITD) and grouping (by F0) in order to estimate a mask for a ‘missing data’ ASR system. A direct comparison of human and machine performance was obtained by using the CASA system as a ‘subject’ in the same computer-based SRT test that was administered to human listeners.

Future work will assess the value of the proposed CASA system as a tool for speech segregation, using a wider variety of speech and noise mixtures that are spatialised using realistic head-related transfer functions. We therefore aim to determine whether a more accurate model of human processing can yield improved performance in engineering applications.

5. Acknowledgements

GJB was funded by EPSRC grant GR/R47400/01. KJP was funded by the EC IHP HOARSE project and the Academy of Finland (grant number 1277811).

6. References

[1] R. F. Lyon, “A computational model of binaural localization and separation,” in *Proc. ICASSP*, 1983, pp. 1148–1151.

[2] N. Roman, D. L. Wang, and G. J. Brown, “Speech segregation based on sound localization,” *J. Acoust. Soc. Am.*, vol. 114, no. 4, pp. 2236–2252, 2003.

[3] 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,” *Speech Comm.*, vol. 43, no. 4, pp. 361–378, 2004.

[4] B. Edmonds, “The role of sound localization in the intelligibility of speech in noise,” Ph.D. dissertation, Cardiff University, 2004.

[5] R. G. Leonard, “A database for speaker-independent digit recognition,” in *Proc. ICASSP*, 1984, pp. 111–114.

[6] I. Ramkissoon, A. Proctor, C. R. Lansing, and R. C. Bilger, “Perceptual segregation of competing speech sounds: the role of spatial location,” *Am. J. Audiol.*, vol. 11, no. 1, pp. 23–28, 2002.

[7] R. Plomp and A. M. Nimpen, “Improving the reliability of testing the speech reception threshold for sentences,” *Acustica*, vol. 18, pp. 43–52, 1979.

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

[9] N. Durlach, “Equalization and cancellation theory of binaural masking level differences,” *J. Acoust. Soc. Am.*, vol. 35, no. 8, pp. 1206–1218, 1963.