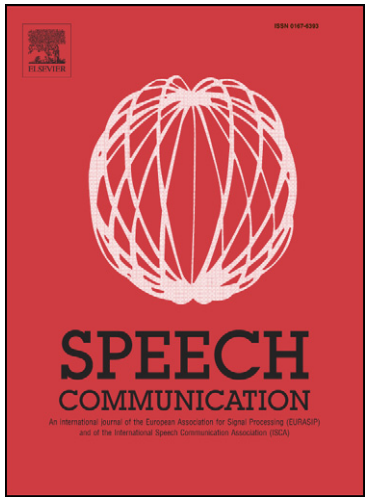
A computational model of binaural speech recognition: Role of across-frequency vs. within-frequency processing and internal noise

Kalle J. Palomäki, Guy J. Brown



A computational model of binaural speech recognition: Role of across-frequency vs. within-frequency processing and internal noise

Kalle J. Palomäki, Guy J. Brown

PII: S0167-6393(11)00049-5
DOI: 10.1016/j.specom.2011.03.005
Reference: SPECOM 1981

To appear in: *Speech Communication*

Received Date: 2 July 2010
Revised Date: 3 March 2011
Accepted Date: 16 March 2011

Please cite this article as: Palomäki, K.J., Brown, G.J., A computational model of binaural speech recognition: Role of across-frequency vs. within-frequency processing and internal noise, *Speech Communication* (2011), doi: 10.1016/j.specom.2011.03.005

A computational model of binaural speech recognition: role of across-frequency vs. within-frequency processing and internal noise

Kalle J. Palomäki[1] and Guy J. Brown[2]

1. Aalto University School of Science and Technology, Department of Computer and Information Science, Adaptive Informatics Research Centre, P.O. Box 15400, FI-00076 Aalto, Finland

kalle.palomaki@tkk.fi

2. Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, United Kingdom

g.brown@dcs.shef.ac.uk

Corresponding author: Kalle J. Palomäki

Tel: +358-9-470-25293

Fax: +358-9-470-23277

Keywords: binaural model, speech recognition, equalization cancellation model, missing data

Abstract

This study describes a model of binaural speech recognition that is tested against psychoacoustic findings on binaural speech intelligibility in noise. It consists of models of the auditory periphery, binaural pathway and recognition of speech from glimpses based on the missing data approach, which allows the speech reception threshold (SRT) of the model and listeners to be compared. The binaural advantage based on differences between the interaural time differences (ITD) of the target and masker is modelled using the equalization-cancellation (EC) mechanism, either independently within each frequency channel or across all channels. The model is tested using a stimulus paradigm in which the target speech and noise interference are split into low- and high-frequency bands, so that the ITD in each band can be varied independently. The match between the model and listener data is quantified by a normalised SRT distance and a correlation metric, which demonstrate a

slightly better match for the within-channel model (SRT: 0.5 dB, correlation: 0.94), than for the across-channel model (SRT: 0.7 dB, correlation: 0.90). However, as the differences between the approaches are small and non-significant, our results suggest that listeners exploit ITD via a mechanism that is neither fully frequency-dependent nor fully frequency-independent.

1. Introduction

Human listeners gain an advantage by listening with two ears when sound impinges upon the head from different directions. In the case of speech signals, intelligibility is improved if the masking signal originates from a direction that is different to that of the target speech (e.g. Hirsh, 1950; Spieth et al., 1954; Hawley et al., 1999). In detection and speech intelligibility experiments, the binaural advantage over monaural listening can be measured in terms of the binaural masking level difference (BMLD) or binaural intelligibility level difference (BILD), respectively. The binaural advantage arises from differences between the interaural time difference (ITD) and interaural level difference (ILD) of the masker and target. For the ILD, this difference originates from monaural listening with the better ear, i.e. the ear in which the signal-to-noise ratio (SNR) is more favourable (Bronkhorst and Plomp, 1988; Edmonds and Culling, 2006). However, the ITD difference between the target and masker yields an unmasking of the target even when the SNR in each ear is equal, which is an advantage that can be achieved only by binaural listening (Bronkhorst and Plomp, 1988). ITDs with a magnitude typical of those that occur in normal binaural listening yield maximal BILDs of between 3-9 dBs, depending on the listening task (Kock, 1950; Schubert 1956; Bronkhorst and Plomp, 1988).

In his pioneering studies Durlach (1963, 1972) suggested that the binaural advantage in hearing can be explained by an equalization-cancellation (EC) model, which is a low-level pre-attentional process. In the equalization ('E') stage of the EC model, the signals originating from one ear (i.e., a mixture of target signal and masking noise) are transformed relative to signals from the other ear, in such a way that the masking noise is equalized and time aligned. The transformed signal is then cancelled by subtraction of the signal from the other ear (the 'C' stage). If the equalization and cancellation processes are done precisely, then the masking noise is eliminated and the target signal remains as a residual (so long as the target differs in interaural time and/or intensity difference from the masking noise). In theory, then, an idealized noise-free EC process can give an infinite improvement in the target/masker ratio. In practice, however, the improvement in target/masker ratio is limited by noise

in neuronal processes.

Durlach (1963) distinguished two types of error that occur at the 'E' stage, and are manifested as noise in the input to the 'C' stage. First, such errors may arise from random jitter in the equalization mechanism. Second, errors may arise due to atypical stimuli that cannot be satisfactorily equalized (e.g., masking signals that have an ITD larger than the maximum possible propagation delay between the two ears). Indeed, noise in analytically formulated models of the EC process has been used to explain psychoacoustic data relating to the BMLD and masking of interaurally delayed stimuli (e.g. Durlach, 1972; van der Heijden and Trahiotis, 1999; Akeroyd, 2004). However, these analytical models make a number of simplifying assumptions. Typically they assume an acoustic input consisting of pure tones and masking noise, which is amenable to analytical study; they do not take sampled audio signals as input. On the other hand, computational models of EC that take simulated auditory-nerve firing patterns as input have been proposed to explain across-frequency independence of the EC-process (Culling and Summerfield, 1995), and various binaural pitch phenomena (Culling et al. 1998; Akeroyd et al., 2001). However, these computational models of EC do not make quantitative predictions of BMLD. Colburn (1973, 1977) proposed a model that was capable of explaining BMLD data. His model is based on detailed neural mechanisms but is unable to process recorded audio signals. Breebaart (2001) also proposed a computational model of EC based on physiologically plausible building blocks. Breebaart's model was tested with psychoacoustic data from BMLD experiments and accepts a time-domain signal so that listener and model responses can be directly compared using the same stimuli. However, the focus of Breebaart's model is signal detection: the 'optimal detector' in the last stage of Breebaart's model performs template matching in order to determine whether a signal is present together with a masker. As a consequence, it is unable to make predictions about the intelligibility of binaural speech signals.

Further insight into the function of the binaural system has come from models of BILD, which represent a step forward from models of BMLD in that they incorporate knowledge of human speech recognition. Beutelmann and Brand (2006) proposed a computational model of binaural speech

perception, in which the BILD was predicted by an EC process followed by estimation of the speech articulation index. Their model achieves a good correspondence with listener data (0.95 correlation). The internal noise models in their work were based on a study by vom Hövel (1984), which presented a revised version of the Durlach EC model.

A separate line of research in binaural modelling has been to build so-called computational auditory scene analysis (CASA) systems (e.g., Lyon, 1983; Roman et al., 2003; Palomäki et al., 2004a), which are motivated by auditory perception but are not intended to replicate psychoacoustic data. These systems often assume that frequency regions that originate from a common azimuthal direction and/or have a common ITD should be grouped. However, psychophysical studies (Culling and Summerfield, 1995; Darwin and Hukin, 1997; Edmonds, 2004; Edmonds and Culling, 2005) suggest that the human auditory system does not use this strategy when segregating concurrent sounds. In an experiment that used double vowel stimuli, Culling and Summerfield (1995) showed that listeners do not use common ITD across frequency as a cue to group vowel formants; subjects were no more likely to group formants lateralized to the same side (i.e., with the same ITD) rather than those lateralized to the opposite side (i.e., with a different ITD). However, in a subsequent study Drennan et al. (2003) argued that the inability of listeners to use ITD related to unnatural ITD cues in the stimuli and the use of synthetic vowel stimuli. With more natural ITDs, ILDs or spatialized stimuli, or when their listeners were trained with the synthetic stimuli, they found binaural advantages based on across-frequency grouping. Using continuous speech, Edmonds (2004) and Edmonds and Culling (2005) investigated the issue of frequency independence in processing of ITD in a speech reception threshold (SRT) test. They conclude that spatial unmasking exploits ITD in a frequency-independent manner (addressed in more detail in Section 2). Their findings are consistent with Culling and Summerfield's (1995) study suggesting that ITD is processed in independent frequency bands, and are incompatible with computational models that use grouping by common ITD. In summary, the psychophysical data show that listeners can exploit a difference in ITD between speech and noise, but that it is not necessary for this difference to be consistent across frequency.

In this study we propose a model of binaural speech recognition and test it against our previously published psychoacoustic data (Brown and Palomäki, 2005), replicating the test (experiment three) by Edmonds and Culling (2005). The main questions addressed in our study are as follows. Firstly, we ask whether the BILD for speech and noise separated by an ITD can be explained using a binaural model based on the EC mechanism, which takes a sampled audio signal as input and recognizes the speech. The output from our model can be scored in exactly the same way as the response of a human listener, and therefore differs from the approach of Beutelmann and Brand (2006); in their model, the BILD was estimated using the speech articulation index without recognizing the speech. Secondly, we ask whether a binaural model that uses ITD in the EC-process independently within each frequency band provides a better fit to listeners' data than one that combines ITD across all frequency bands. Thirdly, we investigate the effects of internal noises in the EC; jitter in neural delay lines and in the equalization gain. To address these questions we propose an approach consisting of a model of the auditory periphery, a binaural processor, and a 'glimpsing' model including an automatic speech recognition (ASR) system that uses the 'missing data' method. The model is compared directly against human performance on the same speech intelligibility tests. The binaural model is based on the EC principle, with its performance limited by internal noise. The glimpsing model finds speech glimpses in which the SNR is favourable for speech. The model applied here is a modified version of Cooke's glimpsing model of human speech recognition (Cooke, 2006). In ASR, glimpses of speech can be exploited by using a 'missing data' approach (Cooke et al., 1994, 2001), in which time-frequency regions are treated as reliable if they consist of relatively noise-free speech (i.e. a speech 'glimpse'), or as unreliable (missing) if they predominantly contain noise. In a comparison of listener data and a glimpsing model a good match was obtained across a number of conditions (Cooke, 2006).

Preliminary versions of this study were published in Brown and Palomäki (2005) and in abstract-only-form in Palomäki and Brown (2008); the current paper describes a substantially improved model that incorporates more refined simulations of the EC process and glimpsing model. The paper is organized as follows. A review of the relevant psychoacoustic studies (Edmonds and Culling, 2005;

Brown and Palomäki, 2005) is given in Section 2. Section 3 describes the computational model, and the model is evaluated in Section 4. We conclude with a discussion in Section 5. The SRT test for comparing machine and human performance used in our previous study (Brown and Palomäki, 2005) is described in the Appendix, which also describes new stimuli employed in the present study.

2. Psychoacoustics background

This section gives a more detailed review of the psychoacoustic study by Edmonds and Culling (2005) and our replication (Brown and Palomäki, 2005) of their experiment. The psychoacoustic test procedure of the Brown and Palomäki (2005) study is described in detail in the Appendix.

*** Figure 1. ***

In order to investigate frequency dependence in the processing of ITD, Edmonds and Culling (2005) conducted a series of three experiments involving an SRT test in which the target speech was split into high- and low-frequency bands and presented with a concurrent speech or Brown noise masker. The Brown noise (a type of noise produced by Brownian motion) used by Edmonds and Culling (2005) was broad-band noise with a 6 dB/octave spectral roll-off, which roughly resembles the spectral shape of speech. Speech material in their SRT test was drawn from the Harvard sentence lists, which consist of English language sentences with a large vocabulary. The review here focuses on two of their tests with Brown noise (experiments one and three), as they are most relevant to the experimental design in the present study. In their experiment one (Figure 1. A and Figure 2. A) the question was asked whether more improvement in speech intelligibility could be achieved by separation of target speech and noise in the full audible frequency band, as opposed to separation only in part of the audible frequency band. The contribution of low- and high-frequency bands to intelligibility was investigated using the stimuli illustrated in Figure 1. A. In the low-contribution condition, the low-frequency bands of the speech and noise had different ITDs, but the high-frequency

bands shared the same ITD. Similarly, in the high-contribution condition, the low-frequency bands of the speech and noise shared the same ITD, but the high-frequency bands had different ITDs. The masking noise in this experiment was always in the centre (zero ITD). The low- and high-contribution results were compared to results obtained with separation in both low- and high- frequency bands ('consistent'), and to a 'same' condition with no ITD separation (all bands of masker and target speech at the centre). The main finding of this experiment was that improvement in intelligibility is larger if both high- and low-frequency bands are separated compared to separation only in the low- or high-frequency band (see Figure 2. A). The improvements were slightly more marked in the low-contribution than in the high-contribution case.

*** Figure 2. ***

Their experiment two was designed to test whether the across-frequency consistency of ITD is used as a cue to group the target speech frequency bands in conditions when the noise is presented in the centre. The results of this condition gave evidence against grouping of the target across frequency. As this experiment is less relevant for the present study, readers are referred to the original publications (Edmonds and Culling, 2005). Finally, their experiment three (Figure 1. B and Figure 2. B) was designed to answer two remaining questions. First, whether across-frequency consistency of the masking noise plays a role. Second, whether different frequency regions of the target speech with different ITDs contribute to binaural unmasking in a simultaneous manner, or whether their impact is pooled over time. As our model (see also Brown and Palomäki, 2005) is based on frequency-independent EC processing, this experiment is the most relevant to the current study. In the experiment, target speech and an interfering sound were split into high and low frequency bands, which were then presented in three ITD configurations (see Figure 1. B). In the 'same-ITD' condition, speech and interference were presented in the same lateral position with the same ITD of +500 μs. In the 'consistent-ITD' condition the target speech was presented with +500 μs ITD and the noise on the

opposite side with -500 μs ITD. In the 'swapped-ITD' condition, the low-frequency band of speech and high-frequency band of noise were presented with +500 μs ITD, and the low-frequency band of noise and high-frequency band of speech with -500 μs ITD. In our previous study (Brown and Palomäki, 2005) we replicated this experiment and compared the human data with the performance of a preliminary computational model. Our experiment (see Appendix for details of the test procedure) differed in three respects from Edmonds and Culling (2005) experiment three: (i) we used spoken digits, to enable the model and listener to be directly compared (ii) we used speech-shaped noise, which is a more effective speech masker than Brown noise (iii) unlike Edmonds and Culling, we did not include competing speech in our experiments. The key findings of both the Edmonds and Culling (2005) original test and our replication (Brown and Palomäki, 2005) are shown in Figure 2 B. The speech intelligibility in the SRT tests was improved in the 'consistent' and 'swapped' conditions when compared to the 'same' condition. However, in both studies the differences in SRT between the 'consistent' and 'swapped' conditions were relatively small (of the two studies, Brown and Palomäki found a slightly larger effect that was statistically significant). Hence, it was concluded that improvements in the SRT were achieved by separation of speech and noise in ITD, but this separation did not need to be consistent across frequency. In both studies this was interpreted as support for frequency-independent processing of ITD in binaural unmasking. Taken together, the results of the experiments reviewed above suggest that ITD-separation contributes to unmasking of the target speech in a way that is frequency independent both in the processing of the target speech (Edmonds and Culling exp. two and three) and the noise masker (Edmonds and Culling exp. three). Therefore, a candidate explanation of the underlying processes can be based on the EC model, which attempts to unmask target speech by removing the noise interference in a frequency independent manner. The present study investigates this possibility.

3. Computational model

The proposed model consists of models of the auditory periphery, binaural processing and glimpsing

model. The latter stage incorporates an ASR system in order to allow direct comparison of the model with human listeners via speech intelligibility tests (see Figure 3.). The binaural processor consists of an EC model, with its performance limited by internal noises that originate from jitter in neural delay lines and equalization gain. The glimpsing model aims to find time-frequency regions ('glimpses') in which the SNR is favourable for speech (Cooke, 2006). The glimpsing model produces a mask in which in each time-frequency region is labelled either as reliable or unreliable evidence for the speech. The acoustic features and mask are then passed to a missing data ASR decoder (Cooke et al., 2001).

In principle, the performance of the model in the SRT test can be tuned by adjusting the internal noise parameters of the EC process and glimpsing model parameters in order to optimise the correspondence with human listeners. In the 'separated' and 'swapped' cases the performance of the model needs to be limited to levels of human performance by addition of a suitable level of internal noise. In the 'same ITD' case the model performance is limited by the lack of interaural cues; the speech recognizer must therefore rely on monaural sound separation in the glimpsing model. In previous studies (Brown and Palomäki, 2005; Palomäki and Brown, 2008) we formulated a glimpsing model that operated blindly on the noisy speech signal only. Here, we achieve near-human performance by making both speech and noise signals available to the glimpsing process (see also Cooke, 2006). This enables the construction of an 'oracle' time-frequency mask, which gives the missing data decoder idealised information about reliable speech regions.

*** Figure 3. ***

3.1. Peripheral model

The auditory periphery is modelled by a bank of auditory (gammatone) filters for each ear, followed by a simplistic model of neuromechanical transduction by inner hair cells. For each ear, $M = 32$ gammatone filters are used, with centre frequencies uniformly spaced between 50 Hz and 8 kHz on an ERB-rate scale. To obtain a crude model of auditory nerve activity, the amplitude output of each gammatone filter is half-wave rectified and compressed by raising it to the power of $\alpha = 0.6$. This is

equivalent to raising the intensity to the power 0.3, and approximates the growth in loudness corresponding to Stevens' Law (Stevens, 1957). The signals are then filtered with a 1 kHz low-pass filter in order to simulate the loss of phase-locking at high frequencies. Finally, to provide acoustic features for the recognizer, the simulated auditory nerve response is sampled at 10 ms intervals and supplemented with delta features. The spectral mean across time was removed using a technique compatible with missing data approaches, as described in our previous publications (Palomäki et al. 2004a, 2004b).

3.2. EC-model

Here the equalization cancellation process is performed in three stages: equalization, estimation of the noise ITD through the cancellation process, and recovery of the speech signal from the noise through cancellation. The EC process was split into three stages in order to implement ITD estimation and models of internal noise in a computationally efficient manner. Separating the ITD estimation and cancellation processes allows the use of longer time windows in the ITD estimation and the use of higher sampling rates for cancellation, which was necessary in order to model jitter in the EC process.

The EC process is performed separately within time-frequency regions, obtained by windowing each channel of the auditory filterbank response into short temporally-overlapping sections. More specifically, each gammatone frequency channel f is split into time frames at 10 ms intervals according to a rectangular window w of length 20 ms. The index t_{fr} is used to refer to the frame index. The output of the EC process is similar to that of the simulated auditory nerve response and can be converted to speech recognition features (see Section 3.1). Signals passed to the EC process are linearized, which is done by expanding the hair cell response by a factor 1/0.6; this completely reverses the compression so that the subtraction process is linear. Following EC, the signals are again compressed according to Stevens' power law. Note that EC was proposed (Durlach, 1963) as a black box model assuming linear input signals. However, in our model the processing remains strictly nonlinear, because the hair cell model introduces nonlinearities (half-wave rectification and low-pass filtering) that are not undone

prior to EC.

3.2.1. *Equalization*

The rms level over the window w is calculated for each time frame t_{fr} for each ear, to yield r_L and r_R, where the subscripts L and R indicate the left and right ears respectively. In order to obtain stable rms estimates over time, r_L and r_R are smoothed by a leaky integrator with a time constant of 95 ms to form $\bar{r}_L$ and $\bar{r}_R$. Linearized left- and right-ear auditory nerve signals x_e are equalized to form $a_e x_e$ as follows

$$a_e x_e(t,t_{fr},f) = x_e(t,t_{fr},f)/\bar{r}_e(t_{fr},f), e \in \{L,R\} \quad (1)$$

where t indexes time at the original sampling rate, which corresponds to the auditory nerve signal temporal resolution.

3.2.2. ITD estimation via cancellation

The cancellation process generates a cancellogram *ecf* from the equalized signals $a_L x_L$ and $a_R x_R$

$$ecf(t_{fr},f,\tau) = \sum_{t=0}^{T-1} | a_L x_L(t,t_{fr},f) - a_R x_R(t+\tau,t_{fr},f) | \quad (2)$$

for each time frame t_{fr}, channel f and time lag τ. The ITD of the noise τ'_n is then estimated by identifying minima in the cancellogram time-frequency bins

$$\tau'_n(t_{fr},f) = \underset{\tau}{\operatorname{argmin}}\left[ecf_{noise}(t_{fr},f,\tau)\right] \quad (3)$$

Furthermore, the ITD estimates τ'_n are accumulated in separate histograms for each frequency channel f over the preceding 500 ms (50 time frames). The histograms are dynamically updated over time to form an online algorithm. The final ITD estimate, $\tau_n(t_{fr},f)$, corresponds to the lag at which there is a maximum in the histogram for each time-frequency bin (t_{fr},f).

3.2.3. Removal of noise in the cancellation process

The performance of the cancellation process in removing acoustic noise is limited by internal noises: jitter in delay lines and equalization gain. The effect of these two types of internal noise was originally postulated by Durlach (1963). Jeffress (1948) proposed that neural delay lines might underlie the

mechanisms of binaural sound localization, and hence similar neural delay mechanisms might also underlie EC. The noise process used here follows the work of vom Hövel (1984) who proposed a modified version of Durlach's model. The previous section presented the ITD estimation, which is here separated from the cancellation process used to perform the actual acoustic noise removal. This is done as the internal noises in the EC process are added after ITD estimation, prior to the actual cancellation process, for signals that are upsampled by factor of 20. In this way the extra computational load caused by upsampling is avoided in the ITD estimation phase. The following formulation describes the EC-residual *d* (upsampled to 20 times the audio sampling rate)

$$d(t,f) = \exp(\varepsilon_L) a_L x_L(t - \tau_n - \delta_L, f) - \exp(\varepsilon_R) a_R x_R(t - \delta_R, f) \tag{4}$$

where $a_L x_L$ and $a_R x_R$ are signals originating from the equalization stage, and the gain errors $\varepsilon_L, \varepsilon_R$ and delay errors δ_L, δ_R are statistically independent Gaussian distributed random variables with zero mean and variances $\sigma_\varepsilon^2 = \sigma_{\varepsilon_L}^2 = \sigma_{\varepsilon_R}^2$ and $\sigma_\delta^2 = \sigma_{\delta_L}^2 = \sigma_{\delta_R}^2$. In this formulation, gain errors ε_L and ε_R are Gaussian in dB units. vom Hövel's model differs in two respects compared to Durlach's model. Firstly, the distribution of gain error is Gaussian in the linear domain in Durlach's formulation, and Gaussian in the logarithmic domain in vom Hövel's. Secondly, in the vom Hövel study both noise in equalization and jitter delay were made dependent on the equalization gain $\alpha = \exp(\frac{a_L}{a_R})$ and cancellation delay Δ according the following formulation

$$\sigma_\varepsilon = \sigma_{\varepsilon_0}\left[1 + \left(\frac{|\alpha|}{\alpha_0}\right)^p\right],\ \sigma_\delta = \sigma_{\delta_0}\left(1 + \frac{|\Delta|}{\Delta_0}\right). \tag{5}$$

vom Hövel used parameters σ_{ε_0} =1.5 dB, α_0=13 dB, *p*=1.6, $\sigma_{\delta 0}$ =65 μs, and Δ_0=1.6 ms. In our study larger variances were chosen (σ_{ε_0} = 3.75 dB and $\sigma_{\delta 0}$ = 150 μs) to provide the best fit to listener data. Note that when linearized input signals are assumed, setting the internal noise parameters to zero would lead to total removal of the noise, leading to an infinite improvement in SNR. In our implementation

the value of temporal jitter $\delta_L(t_{fr}, f)$, $\delta_R(t_{fr}, f)$ and gain errors $\varepsilon_L(t_{fr}, f)$, $\varepsilon_R(t_{fr}, f)$ are varied for each time frame t_{fr} and spectral channel f. The noise (in equation 5) was produced using the Matlab `randn` function, with the exact random sequence held constant across different experimental conditions ("frozen noise") to allow direct comparison of each experiment. The results of the SRT tests using the EC-model were always generated as an average over K randomizations (see Appendix for the SRT test procedure details). In most experiments we set K = 10; in one experiment we set K = 5, as indicated in Section 4.2.

The residual d is divided by the mean of the left $\bar{r}_L$ and right $\bar{r}_R$ rms values, r_{me} to minimize gain variation across time-frequency bins:

$$\hat{d}(t, t_{fr}, f) = \left| d(t, t_{fr}, f) \right| / r_{me} \tag{6}$$

The target signal component in the EC residual includes not only the original target signal, but also a copy of the target signal that is delayed by the target signal ITD and noise cancellation delay, and is 180 degrees phase reversed because of the subtraction process. In the frequency domain this is seen as a comb filter having a notch at zero frequency, which effectively removes the positive slowly-varying energy component of the simulated auditory nerve response. Note that the absolute value is taken in eq. 6 in order to produce the energy envelope required for the ASR system.

3.2.4. Across-channel model

The model described in the previous section presents a 'within-channel' approach to EC, in which the EC process is performed independently in each frequency channel. For comparison, we also consider an 'across-channel' approach that exploits the constancy of ITD across frequency in order to segregate the target speech from the background noise. Across-channel consistency of ITD is a common assumption in 'cocktail-party p rocessors' that use ITD in order to segregate speech from a noisy background (e.g. Lyon, 1983; Roman et al., 2003; Palomäki et al., 2004a).

The difference between the across-channel and within-channel approaches to EC is that the former pools its estimate of the noise ITD over all frequency channels, and uses this single ITD value

in all further steps of the EC process. Hence, instead of separate ITD histograms over each frequency channel (see Section 3.2.2 after eq. 3), the histogram is constructed over all frequency channels. Then the histogram is searched for the peak, which is presumed to be estimate of the ITD of the noise. In Section 4.1.2 we investigate the behaviour of an across-channel scheme.

3.3. Glimpsing model

Here, human speech recognition is modelled using an approach in which glimpses of speech are identified from areas where the SNR is favourable for speech (e.g. Cooke, 2006). In practice, a binary spectrographic mask is used which labels reliable regions with one and unreliable regions with zero (see Figure 4.). Together with spectrograms the masks are then passed to a missing data ASR decoder. In cases where speech and noise are separated by ITD, the glimpsing model is applied and the nerve activity pattern x is obtained after the EC process. This is based on the assumption that finding speech glimpses is a higher-level process, and thus occurs after EC. Therefore, the glimpsing process should operate largely in the same way for both monaural and EC-processed signals. In the monaural channel, the auditory nerve activity is simply summed over left and right ears, which is sufficient when there is no ILD in the stimuli. For signals in which the speech and noise have different ILDs, but the same ITDs, recognition should be based on listening with the better ear (Bronkhorst and Plomp, 1988; Edmonds and Culling, 2006).

In the present study we assume that choosing the information channel, either binaural or monaural (no EC) is idealized. For the stimulus conditions applied in the present study, the model switches between three operational modes to simplify the implementation: (i) monaural, (ii) EC and (iii) mixed monaural-EC. The monaural mode (i) assumes that the signals in all channels propagate through the monaural pathway and are not processed by the EC model. The binaural EC mode (ii) assumes that the signals in all channels propagate through the binaural pathway, and are processed by the EC model. Finally the mixed monaural-EC mode (iii) assumes that the signals in some channels propagate through the monaural pathway and the others through the binaural pathway, with only the

latter being processed by the EC model. In the mixed mode, the switching between monaural and binaural channels is based on *a priori* information. Then in the SRT tests, the modes are used as follows, unless otherwise stated. In the monaural conditions ('same' ITD), the model always uses the monaural channel. In the binaurally separated conditions ('consistent' and 'swapped' ITD) the model chooses either the EC or monaural mode, based on which produces the better recognition accuracy. For the low- and high contribution tests, the model chooses either the mixed monaural-EC mode or the monaural mode, based whichever has the better recognition accuracy. Allowing the model to choose between EC / mixed EC-monaural and monaural mode is based on an assumption of idealized switching, which is similar to choosing the better information channel based on ILD (better-ear listening) or ITD separation (Bronkhorst and Plomp, 1988; Edmonds and Culling, 2006). This improves the recognition in high-SNR cases, when the noise ITD estimate is inaccurate and has been influenced by the speech ITD. Furthermore, the random process in internal noise generation reduces the recognition performance for some particular utterances, which are then recognised more accurately in the monaural mode. This effect is more prominent for low- and high-contribution cases than for the consistent and swapped conditions, in which all frequency bands are separated by ITD. To simplify interpretation of the current modelling results, idealised switching of information channels is assumed – automatic switching was investigated in our previous work (Palomäki and Brown, 2008).

*** Figure 4. ***

Following Cooke (2006), glimpses are identified through an idealized process that uses information from speech and noise signals. The assumption that human listeners have an idealized process for finding glimpses enabled Cooke to obtain a good match to listener data. However, this approach has the obvious drawback that the system is not realizable in practice when the noise signal is not available for comparison, and it does not provide an explanation of how the glimpses are found. An oracle SNR, $SNR_{O,}$ is obtained from the true noise energy $n(t_{fr},f)^2$ by

$$n(t_{fr},f)^2 = \left[x_{s+n}(t_{fr},f) - x_s(t_{fr},f)\right]^2 \tag{7}$$

$$SNR_O(t_{fr},f) = 10\log_{10}\left[\frac{x_S(t_{fr},f)^2}{n(t_{fr},f)^2}\right]$$

where *f* indicates the frequency channel and x_s and x_{s+n} are clean speech (known *a priori*) and noisy speech in the linearized domain, respectively. Depending on whether the operation of the model was in monaural or EC mode, x_{s+n} originated from the peripheral model directly, or through the EC process. If it originated from the EC-model the spectral deviation caused by EC was compensated for before calculating the local SNR. Mask values *m* in each time-frequency bin are set to one if the oracle SNR_O is larger than a threshold θ

$$m(t_{fr},f) = \begin{cases} 1\,, \text{ if } SNR_o > \theta \\ 0\,, \quad \text{otherwise} \end{cases} \tag{8}$$

$$m_g = f(m,S)$$

Then the final glimpsing mask m_g is obtained by processing the original mask *m* to exclude glimpses (i.e., connected regions of reliable components) smaller than a certain minimum size *S*. The glimpsing function *f* is defined as follows: Firstly, regions that are only one time-frequency unit-wide are removed in order to avoid narrow connections between larger areas. In this way, one glimpse that includes a narrow strand between larger areas is divided into two. Secondly, glimpses smaller than a certain minimum size *S* are removed. The minimum glimpse size constraint is based on the assumption that listeners cannot detect very small regions of favourable local SNR (Cooke, 2006). The model we use here is similar to Cooke's (2006) 'glimpses plus background model', in which the 'background model' refers to bounded marginalization (see Section 3.4) rather than pure missing data.

A difference between our model and Cooke's model is in the removal of narrow mask regions. This was added because masks originating from the EC-process have more one-time-frequency unit-wide connections between speech regions than masks produced by the monaural path. These narrow

connections arise from the EC-noise generation process, which varies over each time-frequency bin. This makes SNR variation in adjacent time frequency bins larger than in the masks originating from the monaural path. This problem was addressed both by removing small glimpses, and preventing narrow connections between glimpses in the mask. The restriction on minimum glimpse size has the larger effect, whereas narrow region removal plays a less prominent role. The parameter values for θ and S were sought in a series of recognition experiments. To tune the minimum glimpse size, the principle was to seek a value that was small enough that it did not reduce the monaural system performance, but large enough that it effectively removed small glimpses when the model was in the EC mode. We set S = 45. The final step in the oracle mask parameter adjustment was to seek separate threshold values for the monaural mode (θ = -10 dB) and the EC-mode (θ = -16 dB), which yielded the lowest SRT in recognition experiments when tested across a range of values in 2 dB steps. Note that the θ values are low because the parameter θ is influenced not only by the local SNR but also by other processing that determines the minimum glimpse size (which occurs after the threshold setting in the processing chain). While the mask m (eq. 8) contains all the time frequency units that have local SNR below θ, m_g has considerably less regions labelled as reliable as it contains only those regions in m that are larger than the minimum glimpse size. Considerably higher threshold values (in the neighbourhood of 0 dB) would be required if θ was optimised without the glimpse model. Another observation on the threshold values is the lower threshold for the EC-mode, which is possibly due to fact that the noise estimate obtained by subtraction (eq. 7) does contain some residual speech, whereas in the monaural case the subtraction is made accurately.

3.4. Speech recognizer and data

To compare intelligibility results of human subjects with the proposed model, an ASR system was used. The ASR system consisted of hidden Markov models (HMMs) with state probabilities computed using Gaussian mixture models (GMM) with diagonal covariance. The decoder applied the 'missing data' technique with bounded marginalization for spectrographic features (Cooke et al., 1994, 2001). In

the missing data technique, each spectral feature is labelled as reliable or unreliable. For the reliable features, the observed value is used directly for GMM likelihood estimation. For the unreliable parts, the observation is used only as an upper bound, where the lower bound is set to zero for the spectral features used in this study. The missing data process with bounded marginalization and 'glimpsing' process to identify reliable spectral features is termed the 'glimpse plus background model' in Cooke (2006), which for brevity we refer to as the 'glimpsing' model in the rest of the paper. It is motivated by the phonemic restoration effect, in which the intelligibility of speech containing temporal or spectral gaps is found to improve if the gaps are filled by noise (Warren, 1970; Warren et al., 1997).

In the training phase, acoustic features were computed for the male talkers in the training section of the TIDigits corpus, which contains digits spoken with 22 different American English dialects (Leonard, 1984). These acoustic features were used to train a silence model and eleven word-level HMMs as in our previous study (Palomäki et al., 2004a; Brown and Palomäki, 2005). 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. The test procedure consisted of a digit-based SRT test described in detail in the Appendix. Here we give only a short description. The digit SRT test applied here is a modification of the Edmonds and Culling SRT test that used Harvard sentences. The speech material in our digit SRT test consists of strings of four digits (excluding 'zero' and 'seven' due to their bisyllabicity). The SRT is sought by adaptively adjusting the SNR in steps of 2 dB. The noise and speech in the SRT test are presented in the Edmonds and Culling ITD conditions used in their experiment three ('same', 'consistent' and 'swapped' ITD) and experiment one ('same', 'consistent', 'low-contribution' and 'high-contribution' ITD).

4. Results

4.1. Role of across-frequency vs. within-frequency processing

In this experiment we evaluate the model performance when the ITD estimates are derived

independently within each frequency or by pooling information across frequency. The performances in terms of the SRT are shown in Figure 5. The model was evaluated on the same SRT test that was used in our previous publication (Brown and Palomäki, 2005). The details of the test data are given in the Appendix.

*** Figure 5. ***

4.1.1. Within-channel approach

Edmonds & Culling Experiment 3: Same-, consistent- and swapped-ITD.

Figure 5.A shows a comparison of the model results with human data from the Brown and Palomäki (2005) replication of Edmonds and Culling experiment 3. Note that the same digit-based SRT test was administered to human listeners and the computer model (see Appendix). The axes are aligned so that the average model and listener results share the same vertical centroid while retaining the original scaling (equidistant in dBs). Hence the plot uses different ordinate scales for the model (left) and listeners (right). When the difference between the listener and model SRTs is quantified by the mean over all ITD conditions, the absolute SRT obtained from the model is 0.9 dBs higher on average. Depending on the experimental case, SRTs are from 0.1 to 1.4 dB higher for the model than for listeners. The match between model and listener data is quantified by a normalised SRT distance, obtained by removing the mean SRT across ITD conditions from the model and listener data prior to computing their difference. Based on this distance metric the match is on average 0.5 dBs, and ranges from 0.1 to 0.8 dBs over the ITD conditions.

In the development process the match between the model and listener data was sought in two ways: Firstly, the distance in SRT should be small between ‘consistent’ and ‘swapped’ conditions, as in the listener data. This condition was met by employing a within-channel approach that does not assume across-frequency consistency of ITD. Secondly, the model should approximately replicate the difference in SRT observed for human listeners between the ‘consistent’ / ‘swapped’ ITD conditions, and the ‘same’ ITD conditions. This second condition was met by adjusting the internal noise

parameters. The STDs of jitter in delay lines σ_{ε_0} = 3.75 dB and gain deviation $\sigma_{\delta 0}$ = 150 μs (Section 3.2.3) were set based on a series of tuning experiments so that the difference in dB between the 'same' ITD condition (only monaural cues available) and 'consistent' / 'swapped' ITD conditions (binaural cues available) matched that of human listeners. The difference between the SRT in these conditions gives an estimate of the binaural intelligibility level difference BILD, which was quantified during the parameter adjustment as follows. First, the SRT estimates for the monaural conditions ('same' ITD with splitting frequencies of 750 and 1500 Hz) are averaged. Then the SRT estimates of the binaural conditions ('swapped' and 'consistent' ITD with splitting frequencies of 750 and 1500 Hz) are averaged. Finally the averaged monaural and binaural values were subtracted.

The BILD obtained in this way is 6.2 dB for human listeners and 6.6 dB for the model, giving about a 0.5 dB difference between the machine and human BILD. With the above-mentioned parameter settings of temporal jitter and gain deviation we estimate that the EC process increases the SNR by about 5.0 dB in both the 'consistent' ITD 750 and 1500 Hz conditions, which is 1.6 dBs less than the model BILD.

Statistical analysis conducted on the modelling results using Friedman Anova showed that the ITD condition had a significant effect on the SRT (χ^2 [17,5] = 66.02; $P < 0.01$). Post-hoc examinations using Wilcoxon pairwise tests revealed significant differences in all comparisons in which the 'same' ITD conditions had a higher SRT when compared to either the 'consistent' or 'swapped' ITD ($P < 0.01$). Furthermore, the 'swapped' ITD with splitting frequency of 750 Hz differed significantly from the other conditions in all pairwise comparisons ($P < 0.05$). In comparisons other than those mentioned above, there were no significant differences. In summary, the separation in ITD consistently produced a significant difference while the splitting frequency or swapping the frequency bands did not (with the exception of the 'swapped' ITD 750 Hz splitting frequency condition).

The above-mentioned difference in the absolute SRT between the model and listeners is explained as follows. In the monaural conditions the performance of the model is limited by imperfections that

could originate from any of the monaural processing components, including the feature extraction, glimpsing model and the ASR back-end. The effect of the glimpsing model is further examined in Section 4.3, where we present results with the glimpsing model switched on and off; for a more complete account of the glimpsing model see Cooke (2006). The SRT of the model in binaurally separated conditions was also affected by the limitations of monaural processing in the model, because the model was tuned to match the BILD. Therefore the absolute SRT values were also higher for the model than for listeners in the binaurally separated conditions.

Looking closely at the results for the 'swapped' and 'consistent' conditions, it can be observed that the difference between the SRT in these cases is slightly larger for human listeners than it is for the model. In the model, it is only random variation that could possibly explain any difference between the 'swapped' and 'consistent' ITD cases. It is noteworthy that in their listening tests, Palomäki and Brown (2005) report a significant difference between 'consistent' and 'swapped' ITD conditions while Edmonds and Culling do not. We return to this topic in the discussion.

Edmonds and Culling Experiment 1: low- and high-contribution.

The results presented above demonstrate that the binaural advantage in human listeners can be approximately replicated using the model presented here. However, a possibility remains that the advantage could be due to processing in either the low- or high-frequency bands, but not both. To investigate this possibility we compared human and model performance using the low- and high-contribution stimuli shown in Figure 1A. Model performance was assessed using the digit SRT test (Appendix) and we compare it against the original data from Edmonds and Culling experiment one (which employed a different SRT test). The results are shown in Figure 5.B. To illustrate the match between model and listener data we used different ordinate scales for the model and listener SRTs. The model and listener SRTs are plotted with the same centroid and with an adjusted scaling, which yields an equal difference between the 'same' and 'consistent' ITD conditions. Scaling of the axes was necessary because the digit SRT test amplifies the differences between these conditions, as compared to the SRT test used by Edmonds and Culling. With the adjusted axes settings, the model and listener

SRT patterns are largely similar.

Statistical analysis conducted on the modelling results using the Friedman Anova showed that the ITD condition had a significant effect on the SRT (χ^2 [17,7] = 110.45; $P < 0.01$). Post-hoc examinations using Wilcoxon pairwise tests revealed significant differences across all other conditions ($P < 0.01$) other than in between high- and low-contribution conditions for a splitting frequency of 750 Hz ($P = n.\ s.$). The main observation is that the SRTs for low- and high-contribution stimuli (with binaural cues only on either low or high bands, respectively) were in between those of the 'same' ITD (with only monaural cues), and 'consistent' ITD conditions (with binaural cues on the full bandwidth). This indicates that our model can benefit from binaural separation of both high- and low-frequency bands of the stimuli, as did the listeners in Edmonds and Culling experiment one. A further observation is that the detailed pattern of human and model responses is very similar; the lowest SRTs for both the model and listeners were obtained for the low-contribution ITD stimuli with 1500 Hz splitting frequency, and the highest SRTs for high-contribution stimuli with 1500 Hz splitting frequency.

*** Figure 6. ***

4.1.2. Across channel approach

Figure 6. A shows a comparison of the results of the across-channel model to the listener data in 'consistent' and 'swapped' ITD conditions. The comparison is again made using the same digit SRT test for the model and listeners, with the listener data taken from the Brown and Palomäki (2005) study. The SRT axis is again vertically shifted so that 'consistent' ITD conditions for the model (left) and listener (right) data share the same vertical centroid, while the scaling is unchanged (equidistant in dB). When the difference between the listener and model SRTs is quantified by the mean over all ITD conditions, the absolute SRT obtained from the model is 1.7 dBs higher on average. The results show that the distance in SRT between the 'consistent' and 'swapped' cases is considerably larger for the model than for human listeners. The difference is about twice as much (1500 Hz splitting frequency) or more (750 Hz splitting frequency) when compared to listeners. Using the normalised SRT distance

(defined in Section 4.1.1), the match between the model and listeners is on average 0.7 dBs and ranges from 0.4 to 1.5 dBs.

Statistical analysis on the modelling results using the Friedman Anova again showed that the ITD condition had a significant effect on the SRT (χ^2 [17,5] = 81.24; $P < 0.01$). Post-hoc examinations using Wilcoxon pairwise tests revealed significant differences in all other comparisons ($P < 0.01$) other than between the 'consistent' ITD with 750 Hz and 1500 Hz splitting frequency and between 'same' ITD with 750 Hz and 1500 Hz splitting frequency.

Using the across-channel approach, the results are near to those of the within-channel approach in the consistent condition (difference in SRT less than 0.1 dB), which is the case in which the noise ITD is invariably identified correctly at -500 μs (see Figure 6. A and B). However, considerably higher SRT values are obtained in the swapped condition, when across-channel ITD estimates are from the true noise source only in one band. For the 750 Hz splitting frequency the ITD of the noise is identified based on its high frequency band (ITD of +500 μs), which is explained in this case by the fact that the high frequency band contains more auditory filter bank channels than the low frequency band. This results in a higher proportion of cancellogram minima being allocated to +500 μs in the across-frequency pooled histogram (see Figure 6. B: bottom left panel, Section 3.2.2). This means that the EC-process operates correctly on the high frequency band, but incorrectly on the low frequency band. The SRT result obtained for this case is closest to that of the high-contribution 750 Hz splitting frequency, which is reasonable as both these cases have a binaural advantage in the high frequency band. However, for the splitting frequency of 1500 Hz in the swapped-ITD condition, the noise is allocated both at -500 μs and +500 μs with a slightly higher proportion of estimates at +500 μs (see Figure 6. B: bottom right panel). The 'swapped' ITDs SRT for the 1500 Hz splitting frequency is substantially lower than that for the 750 Hz splitting frequency, and is in fact close to the low-contribution 1500 Hz splitting frequency SRT.

4.1.3. Within vs. across channel model

Sections 4.1.1 and 4.1.2 compare the within- and across-channel models separately against the listener data. In this section we test which model matches better with the Brown and Palomäki (2005) listener data. The correlation coefficient based on normalized zeroth-lag cross-covariance is used as the metric for comparison, because it is not influenced by the mean (absolute SRT) or the variance of the data. We also use a correlation metric that is not influenced by the tuning criterion based on BILD as defined in Section 4.1.1. Specifically, the BILD is normalized to correspond to that of the listener data by subtracting the difference between the listener and machine BILDs from the 'consistent' and 'swapped' condition SRTs. This manipulation was done to remove the effect of the tuning process from both across-channel and within-channel data, because the tuning was performed for the within-channel approach only. Statistically significant correlation coefficients were obtained for both within-channel r^2 = 0.94 ($P < 0.01$) and across-channel models r^2= 0.90 ($P < 0.01$). The difference between the within- and across-channel correlation was investigated using Fisher r-to-z transformation and was found not significant (P = *n. s.*). Correlation coefficients for the BILD normalized data were statistically significant for both the within-channel r^2 = 0.93 ($P < 0.01$) and across-channel models r^2= 0.92 ($P < 0.01$). The difference between the within- and across-channel correlation was found not significant (P = *n. s.*).

In summary, the within-channel model gives a closer match to the listener data compared to the across-channel approach by a small and non-significant margin, which further reduces for the BILD normalized data. Therefore, we cannot regard this result as supporting the within-channel hypothesis of Edmonds and Culling. The difference between consistent and swapped cases is ranked from the smallest to largest as follows: within channel model, listener data and the across channel model. The observation that listener data is ranked in between the within-channel model and the across-channel model suggest a mechanism that is neither fully frequency independent nor fully frequency-dependent.

*** Figure 7. ***

4.2. Effects of internal noise

This section demonstrates the effects of internal noises in the performance of the model using the consistent-ITD stimuli with 750 Hz splitting frequency (see Figure 7.). In this case, to reduce the computational load, we use fewer randomisations (K = 5) when generating internal noise (see Section 3.2.2). Furthermore, the model employed in this section always uses the EC mode (see Section 3.3), instead of choosing the better from the monaural and EC modes. If the monaural mode had been allowed, the effects of internal noises would not have been fully shown in the results – at progressively higher internal noise values, there would be a tendency to select the monaural mode in preference to the EC mode, leading to saturation in the SRT. In order to test the effects of internal noises the STDs of the equalization gain σ_{ε_0} and temporal jitter $\sigma_{\delta 0}$ were separately varied in the neighbourhoods of the values to which they were set (see Section 3.2.3, e.q. 5). When the STD of the temporal jitter $\sigma_{\delta 0}$ is increased from 50 to 250 μs, while the equalization gain deviation is held constant at σ_{ε_0} = 3.75 dB, the SRT increases monotonically from about -17 to -14 dBs. When the equalization gain deviation σ_{ε_0} is adjusted from 2.5 to 5.0 dBs, with the STD of the temporal jitter held constant at $\sigma_{\delta 0}$ = 150 μs, the SRT increases monotonically from about -17 to -13 dBs.

4.3. Evaluation of the glimpsing component of the model

To illustrate the effect of the glimpsing component of the model both in binaural ('consistent' ITD, splitting frequency 750 Hz) and monaural ('same' ITD, splitting frequency 750 Hz) conditions, we conducted a test in which these conditions were compared with and without the glimpsing model. In the condition without the glimpsing model, the missing data processing could in principle be switched off simply by regarding all mask values as reliable (all ones in the mask). Figure 8. shows a comparison of the model with or without missing data processing, which shows a clear advantage due to use of the missing data processing both in monaural (22 dB advantage) and binaural conditions (26 dB advantage). It is also shown that the advantage in performance due to ITD separation weakens in the no-glimpsing model condition. This may relate to the fact that for positive SNRs the ITD estimation of

the noise suffers. If ITD estimation is performed inaccurately, the cues available for separation are mostly monaural.

*** Figure 8. ***

5. Discussion

In this paper we have proposed a computational model for binaural speech recognition and tested it against psychoacoustic data from our replication (Brown and Palomäki, 2005) of an experiment by Edmonds and Culling (2005). The model consists of an equalization-cancellation based binaural model and a monaural model based on recognition of speech from 'glimpses' where the SNR is favourable for speech (see also Cooke, 2006). The EC model was used to remove noise from speech signals and the glimpsing model that includes a mask generation mechanism and a missing data ASR system. The purpose of this work was to evaluate a model of binaural speech recognition, which with certain limitations operates in a similar manner to human listeners, and thus contributes to the development of 'cocktail pa rty processors' a nd a ids in building human-like speech recognition methods. In the development of the model we addressed two main research questions. First, we asked whether within-channel or across-channel processing in EC provides a better model of listeners' SRTs over a number of ITD conditions for speech and noise mixtures. We found that the within-channel model provides a reasonably good fit to listener performance over a range of test conditions including 'same', 'swapped', 'consistent' a s we ll a s 'low–contribution' a nd 'high-contribution' I TD tests. However, there a re differences in the fine detail of model and listener performance. When the within-channel and across-channel approaches are compared, it was noticed that the across-channel model substantially amplifies the small SR T difference be tween 'swapped' a nd 'consistent' I TD observed in the listener performance, while the within channel model predicts almost no difference. Furthermore, correlation between model and listener data was slightly higher for the within-channel model (0.94) than for the across-channel model (0.90), and the difference between correlations further reduced when a BILD-

normalized metric was used: 0.93 for the within-channel approach and 0.92 for the across-channel approach. In neither cases were the differences in correlation between approaches significant. Taken together, these results suggest the existence of an across-frequency mechanism with an efficacy that is in between the purely within-channel and purely across-channel models considered here. Second, we investigated the effect of two kinds of internal noise (temporal jitter and noise in the EC gain) on the ability of the model to replicate listeners' SR Ts. C lose matches were obtained to human data by adjusting the values of these two sources of internal noise.

Results from psychoacoustic tests (Culling Summerfield 1995, Edmonds 2004, Edmonds and Culling, 2005) have suggested that gaining a binaural advantage based on ITD, in recognition or detection, is a frequency-independent process. A subsequent study by Drennan et al. (2003) suggested that across-frequency independence of ITD is a simplification. They showed that after extensive training, listeners could learn to segregate sounds by common ITD. However, large individual differences still remained, and IID cues were found to be much more important than ITD cues.

Our tests indicate that the within-channel model supports Edmonds and Culling's original claim; it predicts little or no difference in the SRT between 'consistent' and 'swapped' ITD conditions. However, the within-channel approach is not able to capture the small but significant differences in the listener SR T between the 'consistent' and 'swapped' ITD cases reported in the study of Brown and Palomäki (2005). This difference is considerably smaller compared to that between the 'consistent' / 'swapped' and the 'same' ITD conditions, which suggests a weak across-frequency dependency. This weak frequency dependency may have remained unnoticed in Edmonds and Culling study (2005), but was observed in the Brown and Palomäki (2005) study that arguably used a more sensitive SRT test. Furthermore, the sensitivity of the Brown and Palomäki test is demonstrated by the fact that it also amplified the differences between the same vs. consistent / swapped –ITD conditions compared to those observed in the Edmonds and Culling study. The most plausible reason for the amplification effects is the different SRT test that we employed, which uses a more effective masker and different speech material (see the Appendix). Further insight into the issue of across-frequency dependency is

gained from the data in Figure 6. in which the across-channel model predicts a considerably larger difference between the 'consistent' and 'swapped' cases.

In Edmonds and Culling (2005; experiment one) the ITD separation of the speech and noise in either the low- or high-frequency bands alone did not explain the binaural advantage, but rather the most advantage was gained when both the high- and low-frequency bands were separated. For experiment one we did not have completely matching data, as our replication (Brown and Palomäki, 2005) of the Edmonds Culling study included only their experiment three. Therefore we compared our modelling results directly to the Edmonds and Culling experiment one data. Largely similar to the Edmonds and Culling study, the performance of the model improved in the cases where the whole frequency band was separated compared to separation in only the low- or high-frequency bands. The difference between the model results and Edmonds and Culling data was that the model amplified the differences. This phenomenon is similar to the effects in the experiment three, which also shows similar amplification in both the model and listener data when they are compared to the Edmonds and Culling results.

Taken together, our findings across a range of ITD conditions and in the within vs. across channel model suggest that there may indeed be some weak across-frequency dependency in the use of ITD. The strength of this effect is in between the within-channel model and the across-channel model. Therefore the model presented here could be improved by incorporating a similar degree of frequency dependency. As suggested by Drennan et al. (2004) learning could be incorporated into our model by mechanisms that perform grouping based on familiar across-frequency patterns of ITD. Such mechanisms were beyond the scope of the current study, but could be addressed in future work.

The model currently has a number of limitations. Firstly, it is functional rather than strictly physiological; for example, feature extraction is based on a highly simplified simulation of the auditory periphery. Also, the key part of the model – the EC-process – is intended to be a 'black box' model rather than one that is physiologically correct. Secondly, following Cooke (2006) the glimpsing model had access to both speech and noise in order to detect speech glimpses. In doing so the models of both

Cooke and the present study produced results that were close to those of human listeners. The obvious drawback of this method is that it does not explain how the glimpses of speech are to be found from the noisy speech only, and thus cannot be used in a practical speech recognition system. Finding glimpses, however, was not the focus of the current study, and the idealized glimpsing process serves the current purpose well.

Previously a number of ways in which glimpses of speech can be found blindly have been suggested (for a review see Barker et al., 2006), e.g. using perceptual criteria such as common harmonicity, estimation of local SNR, or classification of clean speech regions (Setzer et al., 2004). In the two earlier versions of the binaural model presented here we also produced missing data masks using on-line approaches that did not have prior knowledge of the noise. The first one was based on a combined binaural and harmonicity mask (Brown and Palomäki, 2005) and the second based on local noise estimate from speech pauses (Palomäki and Brown, 2008: abstract only). The results of the two above mentioned studies and the present study were collected using the same SRT test (described in the Appendix), so that their results could be directly compared. In the monaural test case ('same' ITD) the results of these systems in terms of SRT, from poorest to the best, were ranked as follows: combined binaural-harmonicity mask (7 dB), mask based on noise estimate from speech pauses (0 dB) and finally the glimpse model in the present study (-9 dB). This indicates that by using *a priori* information about the noise regions in the mask estimation, the machine achieved near-human performance.

In their recent study, Beutelmann and Brand (2006) proposed a model that produces estimates of BILD from audio signals based on the speech articulation index, with a good match to listener data. In the present study the model can also produce estimates of the BILD, but we emphasize that the main goal was not to reproduce accurate BILD estimates, but to study models for binaural processes in connection with an actual speech recognition process, unlike Beutelmann and Brand (2006). For performing only BILD estimation the Beutelmann and Brand model currently has some advantages. In addition to a more concise approach, it does not need to know the noise signal beforehand, which is currently required for our glimpsing model.

According to Durlach's (1963) original formulation, the performance of the EC-process is limited by noise in internal delay estimation and noise in the equalization; a match to listener BMLD data is obtained by tuning these parameters. Durlach's approach was extended by vom Hövel (1984) in order to deal with large ITDs and ILDs, using an approach in which the amount of internal noise was made dependent on ITD or ILD through estimates of the equalization delay or gain. vom Hövel's approach was recently used in estimation of perceived BILD in the study by Beutelmann and Brand (2006). A good match to listener data was obtained with the original STD estimates of vom Hövel in both equalization gain and temporal jitter. Beutelmann and Brand introduced noise in the EC process via a Monte Carlo process, by generating 25 Gaussian random samples for 30 frequency bands, which resulted in 750 replications of the SRT predictions.

In the current study the Monte Carlo process was used with a random sample generated for each time-frequency bin in 10 ms time intervals for 32 frequency channels, with each speech utterance used only once in the speech recognition test. After the EC process this results in 'cleaned' speech spectrograms in which the amount of noise varies randomly in each time-frequency bin. Through using a glimpsing model with a certain minimum glimpse size, these spectrograms were passed to the missing data recognizer. Small glimpses, or narrow patches in glimpses, were excluded to avoid masks with many fragmented small glimpse areas: as the noise varies randomly between each time-frequency bin, it is possible that a fairly noise-free bin can have a very noisy bin next to it (see Sect. 3.3). The noise generation mechanism applied here in connection with the glimpse model led to higher values of the internal noise parameters compared to the vom Hövel and Beutelmann & Brand studies. We sought an explanation for this using an informal small-scale test. Our way of generating noise in the EC process in separate Monte Carlo samples for each time-frequency bins is computationally less expensive compared to, for example, continuous-time and faster/slower temporal variation of noise parameters. However, based on this small-scale experiment, faster smoothly-varying parameter adjustment did not yield substantially different results compared to ones present here. Thus the way in which noise is varied across time-frequency bins is unlikely to explain the relatively larger noise

parameter values compared to those used in the vom Hövel or Beutelmann and Brand studies. Further research is required to clarify the reason for the differences in these internal noise parameter values.

In addition to experiments conducted on Brown noise, Edmonds and Culling made observations using concurrent speech maskers. In the present study, speech maskers were not used in order to simplify our experiments – however, there are no restrictions in the proposed model that would prevent its application to mixtures of concurrent speech. Indeed, signal processing approaches based on a modification of EC have been shown to be effective at cancelling multiple speech maskers (Liu et al., 2001). A complication, however, is that in order to verify the match between our model and listener data using concurrent speech, further psychoacoustic data would need to be collected using the digit SRT test with speech maskers (see Appendix). Using the model to predict SRTs with speech maskers is a future research interest.

To obtain stable ITD estimates for the within-channel model, rather long time windows (500 ms) were used to collect ITD information; this implies that if sources were moving, the performance might not be equally good for the 'consistent' and 'swapped' ITD cases. Again, this is an interesting topic for future investigation. Finally, while listeners were able to take advantage of ITD separation independently of frequency in Edmonds and Culling's (2005) experiments, it appeared that frequency independence did not apply for sounds separated by ILD (Edmonds and Culling, 2006). An interesting future direction would be to model the role of ILD in binaural speech recognition.

Acknowledgement

The work of KJP was supported by the Academy of Finland's support: in project "Auditory approaches to automatic speech recognition" and Adaptive Informatics Research Center. GJB was supported by EPRSRC research grant EP/G009805/1. This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views. The authors wish to thank two anonymous reviewers for their helpful comments.

Appendix: A Speech reception threshold test for psychophysical and computational experiments

In this appendix we describe a SRT test based on spoken digits, which is motivated by the need for a test that is suitable for comparing ASR systems and human listeners. The test was first described in Brown and Palomäki (2005), but is also reviewed here so that the present experiments can be fully understood. The SRT test uses utterances from the TIDigits connected digit corpus (Leonard, 1984), which is a standard corpus for testing noise-robust ASR-algorithms. It has a number of characteristics that suit the current study: well defined training and test sets, highly variable population of speakers, and a small vocabulary size that allows recognition rates closer to humans than larger vocabulary tasks. The use of random digit strings also means that no language models defining word contexts are required for recognition. We note that rather than using spoken digits, Edmonds and Culling (2005) used Harvard sentences spoken by a single speaker.

Ramkissoon et al. (2002) have shown that a SRT test based on a monosyllabic digit vocabulary can be used reliably on human subjects. Similarly, spoken digits have been employed in hearing tests over a range of signal-to-masker ratios (Wilson and Weakley, 2004; McArdle et al., 2005; Wilson et al., 2005). The utterances employed in the SRT test were selected from the TIDigits corpus according to a number of criteria, which aimed to ensure that all trials would be of approximately equal difficulty. Firstly, only four-digit utterances in which each digit contained a single syllable were selected. Hence the digits `oh', `one', 'two', 'three', 'four', 'five', 'six', 'eight' and `nine' were used, whereas `zero' and `seven' were omitted. In addition, 19 male talkers of American English were drawn from 14 dialect groups with the following speaker ID:s: TC, NL, FR, LE, GS, NP, IB, JH, AH, KE, AR, GW, HJ, FG, BN, FT, SA, IP, SL. For further details of the speakers, readers are referred to the TIDigits documentation that also discloses the accents and ages of the speakers. The aforementioned speakers and their utterances were screened by the authors to exclude strongly accented speech, or utterances that exhibited large variations in intensity or fundamental frequency. Informal listening tests, conducted

by the authors, were used to verify that all utterances had approximately equal intelligibility. The sampling rate of the speech signals was 20 kHz.

In the SRT test, utterances were masked by speech-shaped noise that was designed to have the same magnitude response as the long-term spectrum of the TIDigits utterances. Speech babble was generated by summing the speech samples used in the test, then an average magnitude spectrum of the babble noise was generated (FFT size 4096) and a FIR filter of 256 taps was fitted to the magnitude spectrum using the MATLAB `fir2` function (Mathworks, 2008). Finally, Gaussian white noise was passed through the filter to generate speech-shaped noise.

Six lists were constructed for the SRT test from the speech material described above. This was to allow the execution of the test in six experimental conditions (two splitting frequencies and three ITD configurations) all with different speech material. The difficulty of each utterance list could not be balanced by adjusting the initial noise level, because the noise masker was not identical in all conditions. Instead, the sequence of experimental conditions was initially chosen randomly and then rotated for each subject (Brown and Palomäki, 2005; Edmonds, 2004). In the Brown and Palomäki (2005) experiment 12 subjects were measured, hence the lists were rotated twice. Each list consisted of 19 utterances, giving a total of 114 utterances in the test. Within each list, the utterances were produced by 19 different male speakers and the order of speakers was held constant across the lists, which is the same order as in the speaker ID list above.

Our previous psychoacoustic experiment (Brown and Palomäki, 2005) replicated the three ITD-conditions: ‘same’, ‘consistent’ and ‘swapped’ used in Edmonds and Culling (2005) experiment three, which is described in more detail in Section 2 (see also Figure 1. B). To produce these conditions the target speech signal and speech-shaped noise were split into two frequency bands, with the split occurring either at 750 Hz or 1500 Hz. The two bands were separated by a silent gap of one equivalent rectangular bandwidth (ERB; see Glasberg and Moore, 1990), centred on the splitting frequency. In our replication of Edmonds and Culling experiment one, the masker was placed in the centre (Figure 1. A) and did not contain the ERB gap. In the low-contribution stimuli the low-frequency part of speech was

presented with an ITD of +500 μs while the high-frequency part was in the centre (ITD 0). In the high-contribution stimuli the low-frequency part was in the centre and the high-frequency part had an ITD of -500 μs. In the consistent ITD stimulus the masker was in the centre and the target at an ITD of +500 μs.

The noise was presented at a constant level for both ASR and human listeners. In our previous study we used a sound level of 70 dB SPL in the listening test (Brown and Palomäki, 2005). The initial speech level for the adaptive SRT procedure was derived as follows. Prior to the SRT test for each experimental condition, the speech was presented at a level at which it was completely masked by the noise (SNR -26 dB). The speech level was then incremented in steps of 4 dB until the ASR or human subject achieved 50% recognition accuracy and the corresponding SNR was noted. Recognition accuracy was assessed using a standard ASR performance measure defined as 100% – WER, word error rate, which takes into account substitutions, insertions and deletions. This procedure was repeated using two different utterances, and the average SNR was taken as the starting point for the SRT test.

The SRT test itself used an adaptive 1-up / 1-down tracking procedure to adjust the level of the speech (Plomp and Mimpen, 1979). If the ASR system or human subject achieved a recognition accuracy of 75% then the level of the following utterance was reduced by 2 dB, otherwise the level was increased by 2 dB. The SRT was obtained by averaging the SNRs recorded after each level adjustment, with the exclusion of the SNR corresponding to the initial speech level calibration. This gave 17 SRT estimates for each of 6 lists, which totals to 102 for one rotation. The machine performance was estimated based on five or ten rotations in the conditions that involved the EC-model, with five or ten different random seeds in the internal noise generation. The tests on the monaural conditions involved only one rotation as there was no random process taking place. For human subjects the list was rotated twice, so the mean SRT estimates were based on 204 SRT estimates.

The results of the model SRT tests are verified with non-parametric statistical tests, as the data containing discrete decibel readings is not strictly normally distributed. The Friedman Anova is used for testing the main effect and the Wilcoxon-test for post-hoc pairwise comparisons. The data

arrangement was based on the within-speaker averaged SRT, i.e. averaging over six lists (that were in speaker order) and in the case of EC over five or ten randomizations. This results in 17 SRT readings for each ITD-condition. This differs from the statistics used for the listener data in the Brown and Palomäki (2005) study, in which analyses were based on within-listener averaged SRT-readings and MANOVA test procedure.

References

Akeroyd M. A. (2004). "The across frequency independence of equalization of interaural time delay in the equalization-cancellation model of binaural unmasking," J. Acoust. Soc. Am. 116(2), 1135-1148.

Akeroyd M. A., Moore B. C. J. and Moore G. A. (2001). "Melody recognition using three types of dichotic-pitch stimulus," J. Acoust. Soc. Am. 110(3), 1498-1504.

Barker J. P. (2006) "Robust automatic speech recognition" In Wang, D.-L. and Brown, G.J. (Eds.) Computational Auditory Scene Analysis: Principles, Algorithms and Applications Wiley/IEEE Press, 297-350.

Beutelmann R. and Brand T. (2006) "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. 120(1), 331-342.

Breebaart D. J. (2001). "Modeling binaural signal detection," PhD thesis. Technische Universiteit Eindhoven.

Bronkhorst A. W. and Plomp R. (1988). "The effect of head-induced interaural time and level differences on speech intelligibility in noise," J. Acoust. Soc. Am. 83(4), 1508-1516.

Brown G. J. and Palomäki K. J. (2005). "A computational model of the speech reception threshold for laterally separated speech and noise," Proc. Interspeech, Lissabon 4th-8th Sep, 2005, 1753-1756.

Colburn H. S. (1973). "Theory of binaural interaction based on auditory-nerve data. I. General strategy and preliminary results on interaural discrimination," J. Acoust. Soc. Am. 54 (6), 1458-1470.

Colburn H. S. (1977). "Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise," J. Acoust. Soc. Am. 61(2), 525-533.

Cooke M. P. (2006). "A glimpsing model of speech perception in noise," J. Acoust. Soc Am. 119(3), 1562-1573.

Cooke M. P., Green P. and Crawford M. (1994) "Handling missing data in speech recognition," Int. Conf. Spoken Lang. Proc., 1555-1558.

Cooke M. P., Green P., Josifovski L. and Vizinho A. (2001). "Robust automatic speech recognition with missing and unreliable acoustic data," Speech Comm. 34, 267–285.

Culling J. F. and Summerfield Q. (1995). "Perceptual separation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay," J. Acoust. Soc. Am. 98(2), 785-797.

Culling J. F., Summerfield A. Q. and Marshall D. H. (1998). "Dichotic pitches as illusions of binaural unmasking I. Huggins' pitch and the 'binaural edge pitch'," J. Acoust. Soc. Am. 103(6), 3509-3526.

Darwin C. J. and Hukin R. W. (1997). "Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity," J. Acoust. Soc. Am. 102(4), 2316-2324.

Darwin C. J. and Hukin R. W. (2000). "Effectiveness of spatial cues, prosody, and talker characteristics in selective attention," J. Acoust. Soc. Am. 107(2), 970-977.

Drennan W. R., Gatehouse S. and Lever C. (2003). "Perceptual segregation of competing speech sounds: the role of spatial location," J. Acoust. Soc. Am. 114 (4), 2178-2189.

Durlach N. I. (1963). "Equalization and cancellation theory of binaural masking-level differences," J. Acoust. Soc. Am. 35(8), 1206–1218.

Durlach N I. (l972). "Binaural signal detection: equalization and cancellation theory," Foundations of Modern Auditory Theory, Editor: J. V. Tobias, Academic Press, New York. Volume 2, 371-462.

Edmonds B. (2004). "The role of sound localization in the intelligibility of speech in noise," Ph.D.

dissertation, Cardiff University, 2004.

Edmonds B. A. and Culling J. F. (2005). "The spatial unmasking of speech: evidence for within-channel processing of interaural time delay," J. Acoust. Soc. Am. 117(5), 3069-3078.

Edmonds B. A. and Culling J. F. (2006). "The spatial unmasking of speech: evidence for better-ear listening," J. Acoust. Soc. Am. 120(3), 1539-1545.

Glasberg B. R. and Moore B. C. J. (1990). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. 47, 103-138.

Hawley M. L., Litovsky R. Y. and Colburn H. S. (1999). "Speech intelligibility and localization in a multi-source environment," J. Acoust. Soc. Am. 105 (6), 3436- 3448..

Hirsh I. J. (1950). "The relation between localization and intelligibility," J. Acoust. Soc. Am. 22(2), 196-200.

Jeffress L. A. (1948). "A place theory of sound localization," J. Comparat. Physiol. Psychol. 41, 35-39.

Kock W. E. (1950) "Binaural localization and masking," J. Acoust. Soc. Am. 22(6), 801-804.

Leonard R. G. (1984). "A database for speaker-independent digit recognition," in Proc. Int. Conf. Acoust. Speech Sig. Proc., 1984, 111–114.

Liu C., Wheeler B. C., O'Brien W. D. Lansing C. R., Bilger R. C., Jones D. L. and Feng, A. S. (2001) A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers. J. Acoust. Soc. Am., 110 (6), 3218-3231.

Lyon R. F. (1983). "A computational model of binaural localization and separation," in Proc. Int. Conf. Acoust. Speech Sig. Proc., 1983, 1148–1151.

Mathworks (2008). MATLAB. http://www.mathworks.com

McArdle R. A., Wilson R. H., and Burks C. A. (2005). "Speech recognition in multitalker babble using digits, words, and sentences," J. Am. Acad. Audiology 16, 726-739.

Palomäki K. J. and Brown G. J. (2008) A computational model of binaural speech intelligibility level difference, J. Acoust. Soc. Am. 123, 3715.

Palomäki K. J., Brown G. J. and Wang D. L. (2004a). "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," Speech Comm. 43(4), 361–378.

Palomäki K. J., Brown G. J. and Barker J. P. (2004b). " Techniques for handling convolutional distortion with "missing data" automatic speech recognition," Speech Comm. 43(1-2), 123-142.

Plomp R., and Mimpen A. M. (1979). "Improving the reliability of testing the speech-reception threshold for sentences," Audiology 18, 43–52.

Ramkissoon I., Proctor A., Lansing C. R., and Bilger R. C. (2002). " Digit speech recognition thresholds (SRT) for non-native speakers of English," Am. J. Audiol. 11, 23–28.

Roman N., Wang D. L. and Brown G. J. (2003). "Speech segregation based on sound localization," J. Acoust. Soc. Am. 114(4), 2236–2252.

Schubert E. (1956) "Some preliminary experiments on binaural time delay and intelligibility", J. Acoust. Soc. Am. 28(5), 895-901

Spieth W., Curtis J. F. and Webster J. C. (1954). "Responding to one of two simultaneous messages," J. Acoust. Soc. Am. 26 (3), 391-396.

Stevens S. S. (1957). "On the psychophysical law," Psychological Review 64(3),153–181.

van der Heijden M. and Trahiotis C. (1999). "Masking with interaurally delayed stimuli: The use of "internal" delays in binaural detection," J. Acoust. Soc. Am. 105(1), 388-399.

vom Hövel H. (1984) Zur bedeutung der ubertragungseigenschaften des aussenohrs sowie des binauralen hörsystems bei gestörter sprachubertragung", Dissertation, Fakultät fyr Elektrotechnik, der Rheinisch-Westfälischen Technischen Hochschule, Aachen.

Warren R. M. (1970) "Perceptual restoration of missing speech sounds," Science 167, 392-393.

Warren R. M., Hainsworth K. R., Brubaker B. S., Bashford JR. J. A., Healy E. W. (1997) "Spectral restoration of speech: Intelligibility is increased by inserting noise in spectral gaps," Perception and Psychophysics, 59(2), 275-283.

Wilson R. H., and Weakley D. G. (2004). "The use of digit triplets to evaluate word-recognition abilities in multitalker babble," Seminars in Hearing 25(1), 93-111.

Wilson R. H., Burks C. A., and Weakley D. G. (2005). "A comparison of word-recognition abilities assessed with digit pairs and digit triplets in multitalker babble," Journal of Rehabilitation Research and Development 42(4), 499-510.

Figure legends

Figure 1. Schematic diagrams of the stimuli. A. Stimuli used in Edmonds and Culling experiment one showing the same, consistent and low- and high-contribution stimuli. B. Edmonds and Culling experiment three and our replication of it (Brown and Palomäki, 2005) showing the same-, consistent- and swapped-ITD configurations.

Figure 2. Results of the psychoacoustic experiments by Edmonds and Culling (2005) and Brown and Palomäki (2005). A. SRTs for Edmonds and Culling (2005) experiment one with Brown noise interference. Conditions are denoted as same- (○), consistent-ITD (□) low- (∇) and high-contribution (Δ). B. Speech reception thresholds (SRTs) are shown for Brown and Palomäki (2005) (solid lines and closed symbols) and for Edmonds and Culling experiment three (dashed line and open symbols) from which only the Brown noise case is shown here. Conditions are denoted as same- (●, ○), consistent- (■, □) and swapped-ITD (♦, ◊). Open symbols denote the Edmonds and Culling data, and closed symbols denote the Brown and Palomäki data.

Figure 3. Schematic diagram of the model. The major components are peripheral model, equalization-cancellation (EC) processing and glimpsing model including automatic speech recognition (ASR).

Figure 4. Demonstration of mask generation in the glimpsing model, showing auditory spectrograms and a mask derived from the glimpsing model for the u tterance "6185" for 750 Hz splitting frequency stimulus: A. In clean conditions, B. mixed with speech-shaped noise (SNR=-9 dB) and C. the corresponding time-frequency mask, in which reliable speech regions are shown in black.

Figure 5. Comparison of listener data and within-channel model in the SRT task. In each graph the left and right Y-axis scales are for the computer model and the listener SRT, respectively. A. SRTs for

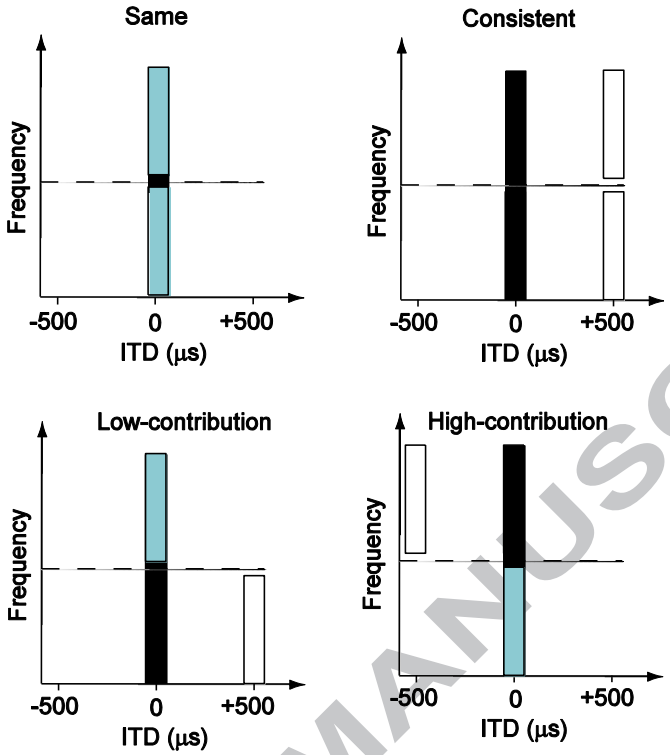
the model and human listeners from the Brown and Palomäki (2005) study based on Edmonds and Culling experiment 3. The conditions shown are same- (●, ○), consistent- (■, □) and swapped- (♦, ◊). B. SRTs for the within-channel model and listeners from Edmonds and Culling experiment 1 are shown for the same- (●, ○) high-contribution- (▲, Δ) low-contribution- (▼, ∇) and consistent- (■, □) conditions. Error bars indicate standard error of the mean (SEM). Closed symbols denote the model results and open symbols denote the listener results.

Figure 6. Comparison of listener data and across-channel model in the SRT task. A. SRTs for the across-channel model and human listeners from the Brown and Palomäki (2005) study, for the same- (●, ○) consistent- (■, □) and swapped- (♦, ◊) ITD-conditions. Error bars indicate the SEM. Closed symbols denote the model and open symbols denote the listener results. B. Histograms for ITD estimates using the across-channel model for swapped- and consistent-ITD conditions.

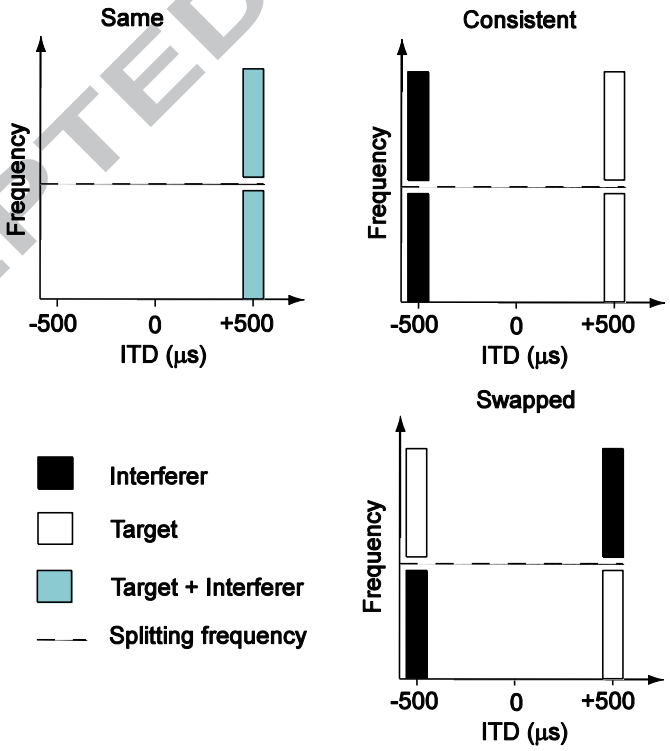
Figure 7. The effect of A. temporal jitter noise in delay lines and B. noise in the equalization process on the SRT of the computer model. Error bars indicate the SEM.

Figure 8. Comparison of the SRT of the computer model with and without missing data-processing for 750 Hz splitting frequency stimuli, for the same- (●) and consistent-ITD (■) configurations. Error bars are negligible on the plotting scale.

Figure 1.

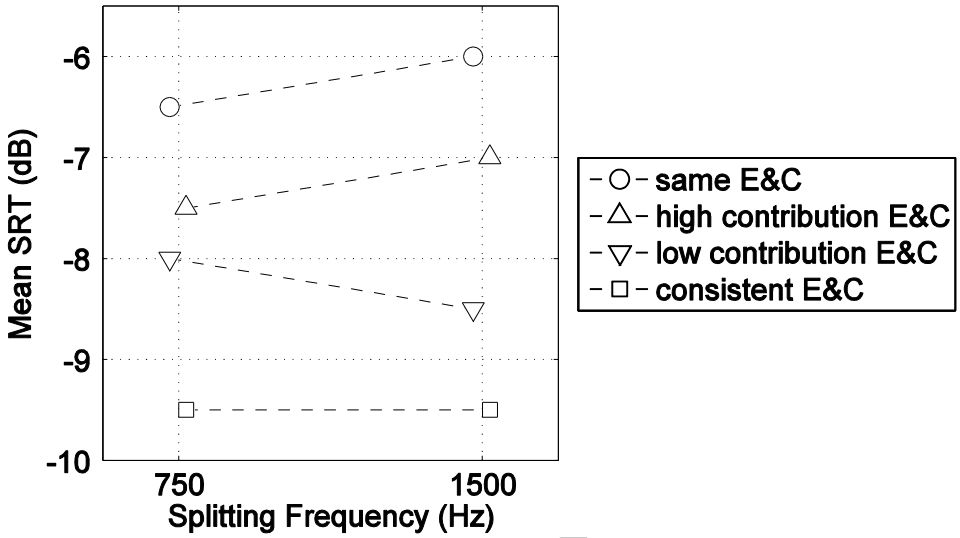


A.

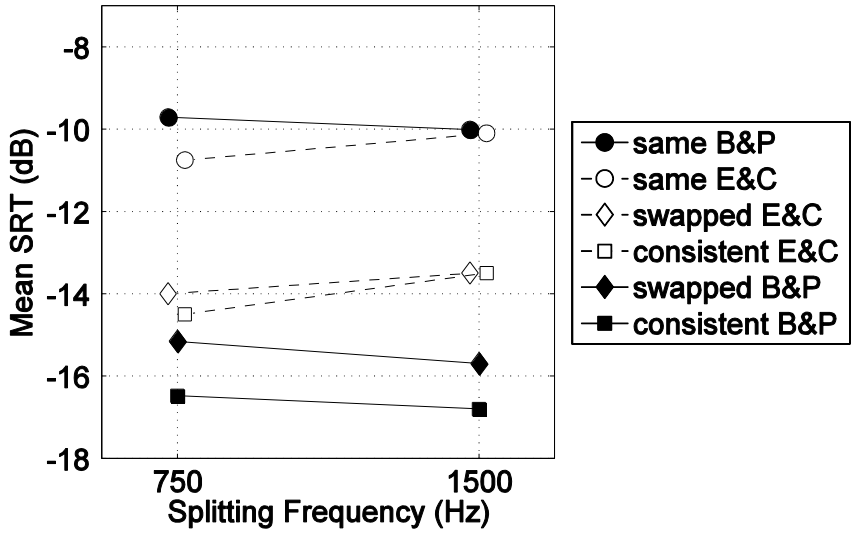


B.

Figure 2.



A.



B.

Figure 3.

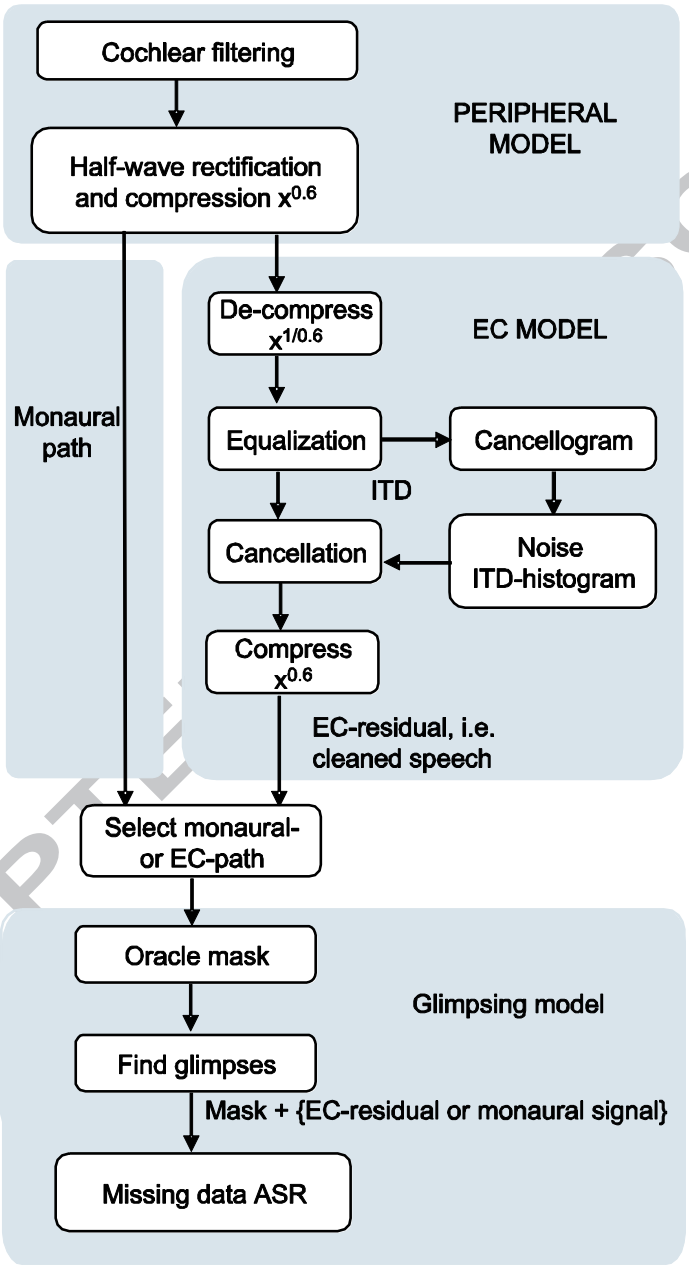


Figure 4.

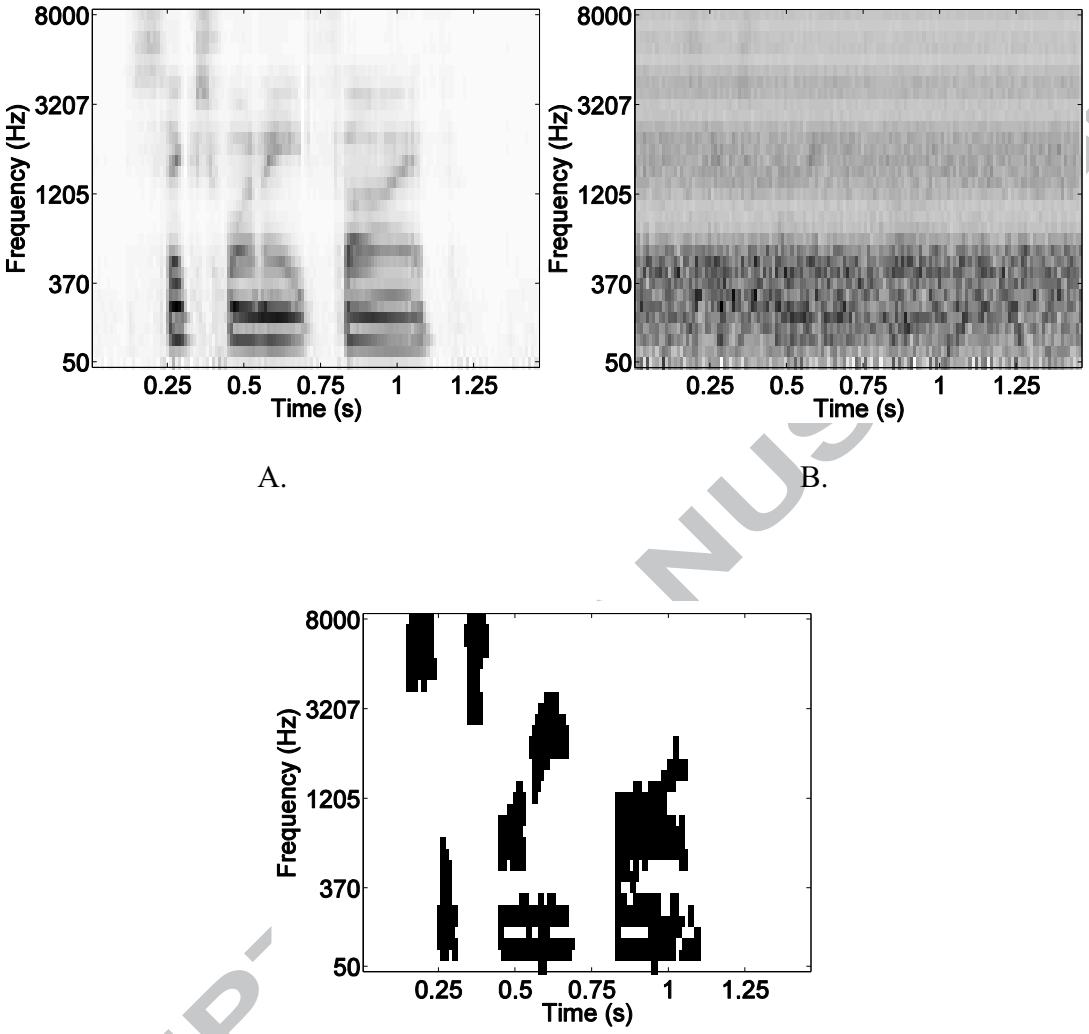
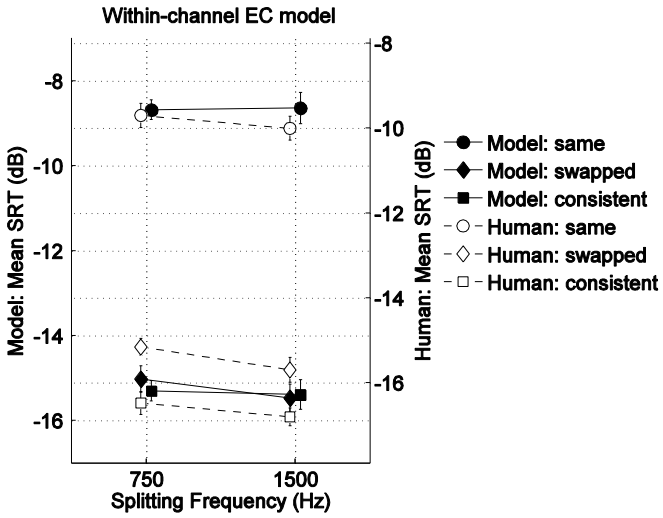
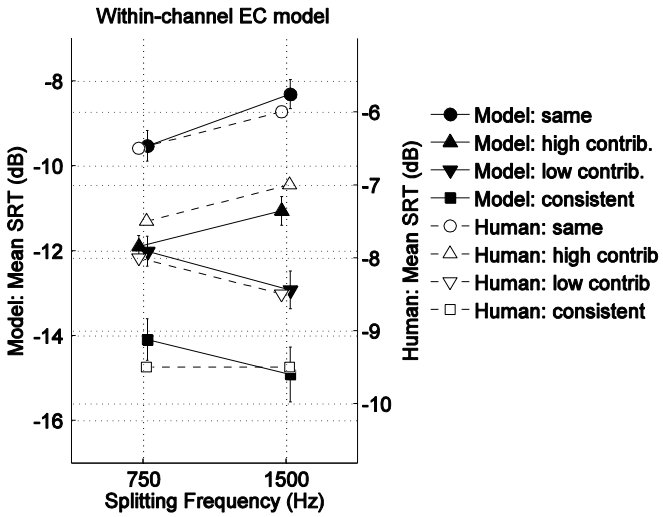


Figure 5.

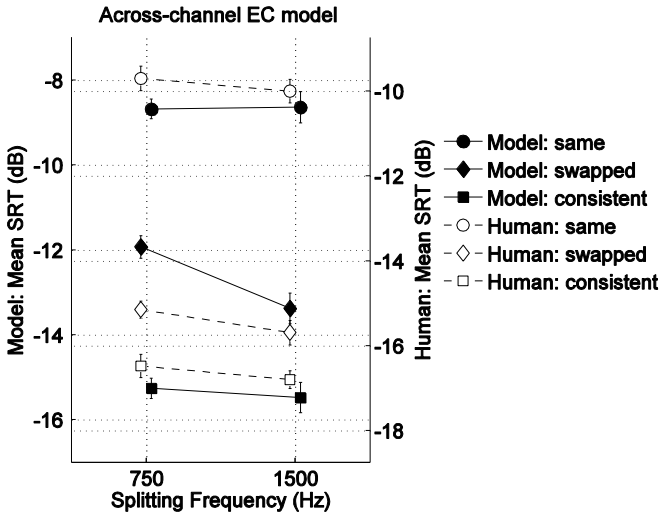


A.

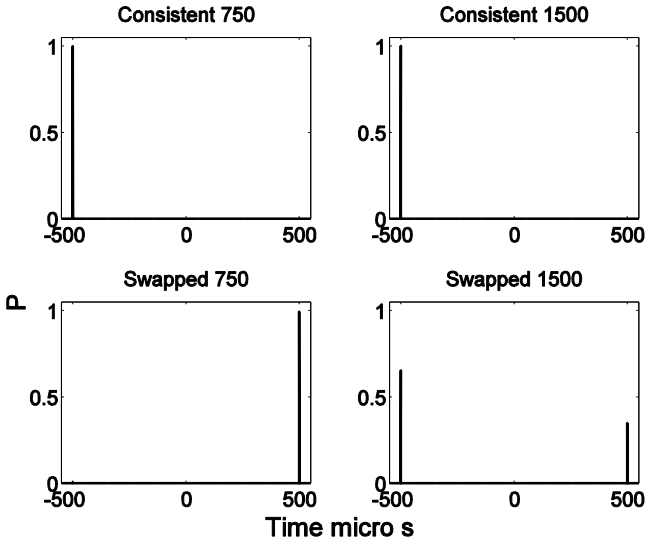


B.

Figure 6.

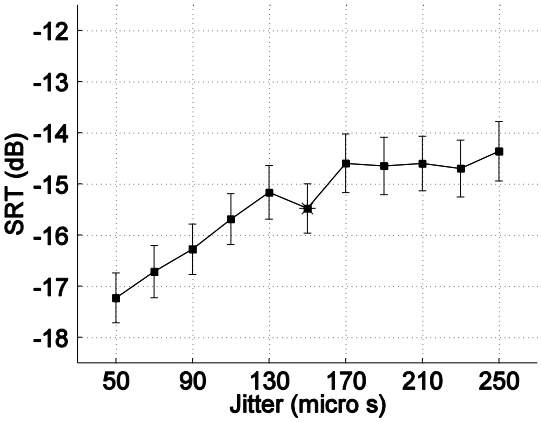


A.

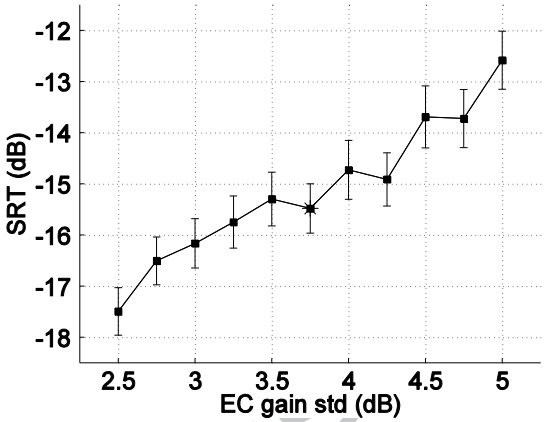


B.

Figure 7.



A.



B.

Figure 8.