Binaural intelligibility prediction based on the speech transmission index[a)]

Sander J. van Wijngaarden and Rob Drullman
TNO Human Factors, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands

(Received 9 February 2007; revised 14 March 2008; accepted 14 March 2008)

Although the speech transmission index (STI) is a well-accepted and standardized method for objective prediction of speech intelligibility in a wide range of environments and applications, it is essentially a monaural model. Advantages of binaural hearing in speech intelligibility are disregarded. In specific conditions, this leads to considerable mismatches between subjective intelligibility and the STI. A binaural version of the STI was developed based on interaural cross correlograms, which shows a considerably improved correspondence with subjective intelligibility in dichotic listening conditions. The new binaural STI is designed to be a relatively simple model, which adds only few parameters to the original standardized STI and changes none of the existing model parameters. For monaural conditions, the outcome is identical to the standardized STI. The new model was validated on a set of 39 dichotic listening conditions, featuring anechoic, classroom, listening room, and strongly echoic environments. For these 39 conditions, speech intelligibility [consonant-vowel-consonant (CVC) word score] and binaural STI were measured. On the basis of these conditions, the relation between binaural STI and CVC word scores closely matches the STI reference curve (standardized relation between STI and CVC word score) for monaural listening. A better-ear STI appears to perform quite well in relation to the binaural STI model; the monaural STI performs poorly in these cases. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2905245]

PACS number(s): 43.71.An, 43.66.Pn [DOS] Pages: 4514–4523

I. INTRODUCTION

Speech intelligibility is most accurately and representatively measured by using subjective test procedures, involving panels of human test subjects. Unfortunately, subjective tests are cumbersome and expensive. For this reason, researchers, engineers, and acoustics consultants often rely on objective procedures to predict speech intelligibility. Examples of such procedures are the articulation index (Kryter, 1962), the speech intelligibility index (SII) (ANSI, 1997), and the speech transmission index (STI) (IEC, 2003; Steeneken and Houtgast, 1980). The SII and the STI are considered to represent the state of the art in intelligibility prediction. Although these models are generally successful in predicting intelligibility across a wide range of conditions, there are always conditions for which inaccurate results are obtained. An important source of prediction errors is that fact that standardized versions of the STI and SII are monaural models; they are based on single-channel (or single ear) estimates. By extending the prediction models to cover aspects of binaural hearing, their scope is extended to applications for which otherwise inaccurate results would be obtained.

This paper describes an extension of the STI model to a binaural intelligibility prediction model by adding algorithms that simulate binaural interaction. A similar approach could probably be adopted to modify the SII. The current paper focuses only on the STI largely for practical reasons: the STI is more widely used by acoustic consultants and engineers, due to the availability of measuring devices that are capable of rapidly producing STI values, through direct measurements.

The STI was originally designed to predict intelligibility in diotic listening conditions based on measurements with a single microphone. It is beyond the scope of this paper to give a complete description of the STI method. Basically, the method assumes that the intelligibility of a transmitted speech signal is related to the preservation of the original spectrotemporal differences between the speech sounds. These spectral differences may be reduced by bandpass limiting, masking noise, nonlinear distortion components, and distortion in the time domain (echoes and reverberation). The reduction of these spectral differences can be quantified by looking at the modulation transfer in a number of frequency (octave) bands. More background information on the STI can be found in the literature (e.g., Steeneken and Houtgast, 1980; IEC, 2003). Given the diotic listening conditions of the traditional STI, this means that all binaural (dichotic) intelligibility benefits are disregarded. The resulting inaccuracy may be considerable if sources of speech and interfering noise are separated spatially. Intelligibility, and hence the STI, depends on the relative positions of source (speech/noise) and listener within a certain space.

Potentially, the extension of the STI to a binaural model could reduce the general applicability; changes to the model might affect its validity in other conditions, unless specific precautions are taken. Care must be taken to ensure that the measured STI value is unchanged compared to the original model if binaural hearing is presumed not to play any role.

[a)]Part of this work was presented at the 151st ASA meeting in Providence, RI, 5–9 June 2006.

 0001-4966/2008/123(6)/4514/10/$23.00

Also, the attractive features of the STI method should be kept intact. In summary, this leads to the following requirements for the development of a binaural STI method:

(a) fast (15 s) measurements with a test signal in any environment;
(b) representative results in noise, reverberation, and nonlinear distortion;
(c) simple model, with very few model parameters, none of which are "tuned" to any specific application;
(d) feasible as an extension to current STI measuring devices.

These requirements almost naturally lead to the conclusion that a good option is to develop a model extension that allows STI measurements in the same way as currently standardized, but with two microphones (or rather an artificial head) instead of one. This should be achieved by incorporating a model of binaural listening into the STI framework.

II. BINAURAL INTELLIGIBILITY MODELLING

A. Background

The benefit of listening to speech with two ears instead of one in conditions with background babble is known as the cocktail party effect (Cherry, 1953). A significant body of scientific research on this topic (Bronkhorst, 2000), spanning half a century, provides ample resources to draw from for devising binaural intelligibility models.

Binaural speech intelligibility tends to be better than monaural intelligibility because of the contributions of two factors: head shadow and binaural interaction. Head shadow may result in an (effective) speech-to-noise ratio that is better at one ear than the other; by using the "better-ear" signal, the intelligibility is improved. This effect, based on interaural level differences, can probably be incorporated in the STI model relatively easily by using separate measurements corresponding to the left and right ears. The main question is how to choose from both ears: perhaps by selecting the best overall STI or selecting the best signal on a band-by-band basis (cf. Edmonds and Culling, 2006).

The effect of binaural interaction on speech intelligibility is primarily related to interaural time differences, although interaural decorrelation may also play a role in more reverberant environments (Bronkhorst and Plomp, 1990). The literature presents various models of binaural interaction (e.g., Stern and Trahiotis, 1995), mostly based on the concept of binaural cross correlation (Jeffress, 1948; Zwicker and Henning, 1985; Raatgever and Bilsen, 1986). Cross-correlation models of binaural processing help explain various auditory phenomena related to binaural hearing, such as lateralization, binaural pitch, and binaural masking level differences, while also appearing physiologically feasible (Colburn, 1995). Models of binaural interaction have been refined to a level at which detailed predictions can be obtained for many phenomena. The most important of these models, which are powerful but also quite complex, are the equalization-cancellation model (Durlach, 1963, 1972) and the auditory-nerve-based model (Colburn, 1973).

Interaural time differences (ITDs) and interaural level differences (ILDs) both contribute to an improvement in intelligibility over monaural listening. However, these contributions are not mutually independent. In an anechoic environment, an improvement in the signal-to-noise ratio (SNR) corresponding to 50% sentence intelligibility of up to 8 dB was found due to ILDs, while the improvement due to ITDs was up to 5 dB. However, the combined effect was at most 10 dB (Bronkhorst and Plomp, 1988).

A quantative model for predicting binaural advantages and directional effects in speech intelligibility was presented by Zurek (1993). It models speech and interference in 1/3-octave bands, accounting for the binaural interaction by using interaural level and phase differences. Zurek's model proved to give reasonably adequate predictions of existing data in a number of spatial configurations. However, the model is restricted to include masked speech in an anechoic environment only. Reverberation (for both speech and interference) is not incorporated, which makes this model not very suitable for typical STI applications (Houtgast and Steeneken, 2002).

Recently, Beutelmann and Brand (2006) presented a binaural intelligibility prediction model based on an extended equalization-cancellation process and the SII. They used three different acoustic environments (anechoic, office room, and cafeteria) to measure the speech reception threshold (SRT) with normal-hearing and hearing-impaired listeners. The overall correlation between predicted and observed SRTs proved to be quite high (0.95). Although, in principle, capable of handling reverberation, their model was only tested for near-field speech. Beutelmann and Brand (2006) proposed to use the STI instead of or as a correction on the SII. However, their binaural processing is quite complex, which we consider a drawback for application with the STI.

B. Incorporating binaural effects in the STI

Over the past decades, the STI model has gradually evolved from a very simple procedure suitable for a limited set of applications to a widely applicable model that is representative for most practical situations in which speech communication occurs. Features have been added to the model. For example, the current version of the STI incorporates the effects of mutual dependence between frequency bands and also the dependence of auditory masking curves on the absolute level. Whenever the model was enhanced or modified, care was taken to adhere to the following principles.

(a) The relation between STI and subjective intelligibility must remain unchanged after modification (i.e., the new version of the STI must exactly replicate results obtained with past versions, except in those cases where the "old" model was proven inaccurate).
(b) The model parameters of the STI are never tuned to a specific application. There is but one universal set of STI model parameters.
(c) STI improvements are always aimed at improving the accuracy for certain conditions. However, this always

makes the model more complex. The added complexity of a model modification must be proportional to the achieved accuracy improvement.

By sticking to the principles given above, the STI model has over the last years improved significantly without losing touch with engineers and consultants who already used it. Especially, the last principle on the list has turned out to be of great importance for the standardization of the STI. Not all modifications to the STI, proposed in the literature, have therefore been incorporated into the International Electrotechnical Commission (IEC) standard. If a new addition to the model doubles or triples its complexity, this will clearly affect the cost of STI measuring equipment. The increase in performance should warrant such an increase in cost.

For our intended extension of the STI to binaural listening conditions, model complexity is a realistic concern. The use of a comprehensive state-of-the-art binaural interaction model would greatly increase the complexity of the entire STI model. Our conclusion is that we need to look for a simplified quantification of the effects of binaural interaction. This will be less general and probably less accurate than the state of the art in binaural modeling. However, the aim is not to minimize the resulting prediction error—just to reduce this error to the same order of magnitude as other sources of variance in the STI model. Greater accuracy of the binaural interaction model would be meaningless since the overall error in the STI would then be determined by other factors (Houtgast *et al.*, 1980).

Our current proposal is to incorporate binaural interaction through the estimation of a simple interaural cross correlogram. In this, we follow the approach by Jeffress (1948), which assumes a mechanism fundamentally related to cross correlation. It is customary to incorporate auditory filter band models and hair-cell models in such an estimation of the cross correlogram. Our current aim is to simplify this as far as possible. The basic idea is visualized in Fig. 1, which shows the way in which interaural correlograms could be represented in the context of the STI model. Signals corresponding to the left (L) and right (R) ear are measured and divided into octave bands (centered from 125 Hz to 8 kHz), as customary in the STI model. In each octave band (or at least the ones covering the approximate frequency range in which humans can analyze interaural time relations, presented in gray in Fig. 1), the interaural cross correlation is calculated. The signal is reconstructed at several "internal" time delays of up to (plus or minus) a few milliseconds. Next, these internal spectral representations are analyzed in the usual way, as if corresponding to a single-channel STI measurement. This yields a quantification of the internal modulation transfer for each octave band at each interaural delay time.

The final problem is to select the most representative internal delay time for each octave band. Assuming that human binaural processing results in a straightforward strategy of intelligibility optimization, the most likely candidate is the internal delay at which the maximum modulation transfer is observed. By using these results, an overall (binaural) STI can be calculated.

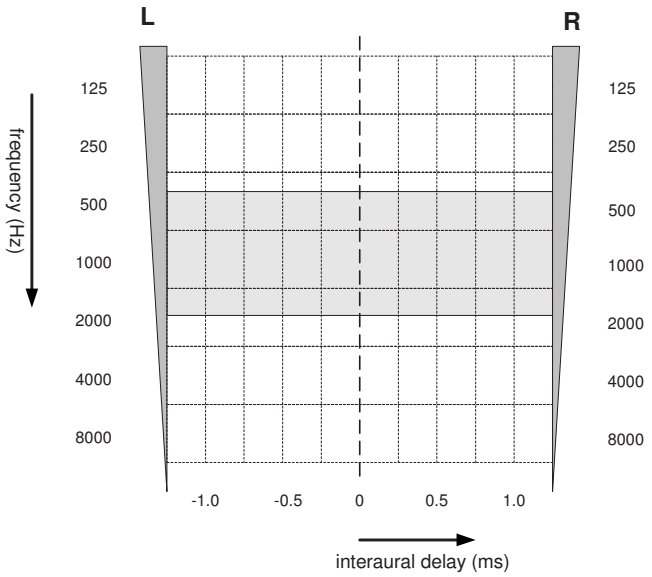

FIG. 1. Visualization of a "grid" for displaying interaural cross correlograms in the context of the STI model. Left (L) and right (R) ear signals are divided into octave bands. In applicable octave bands (gray rectangle), the interaural cross correlation is calculated and the signal is reconstructed at several "internal" time delays.

Within the framework described here, a number of different binaural STI implementations can be thought of. Construction and implementation of such a model comes down to choosing which degree of simplification is accepted and choosing model parameters. This process is outlined in the next section.

III. IMPLEMENTATION OF A BINAURAL STI MODEL: DESIGN CHOICES

A binaural STI model based on the framework described in the previous chapter was designed and implemented in MATLAB®. The measurements on which binaural STI calculations are based are straightforward extensions of the normal standardized STI measurements with the following adaptations:

(a) Each binaural STI measurement is based on a two-channel recording, obtained using an artificial head. This artificial head marks the position of the (simulated) listener.
(b) The test signal—played back at the position of a simulated talker—can be any standardized telecommunications (STI) signal, such as STITEL, STIPA, or full STI (cf. IEC, 2003). The exception is room acoustics (RASTI) , which cannot be used since the 1 kHz band, which is essential in binaural listening, is not included in the signal

A first approximation of binaural speech intelligibility is obtained by calculating the STI from the left and right ears of the artificial head separately. This can be done by using a standard equipment. By taking the highest STI (better ear), the effects of interaural level differences are taken into account. This approach can be slightly refined by taking the best signal (left-right) on a band-by-band basis. The better-

ear approximation is expected to underestimate the contribution of frequency bands in which interaural time differences can be used to (perceptually) enhance the speech signal. This is where the model could be extended.

Since the frequency range in which the most useful binaural interaction for speech intelligibility takes place extends from 500 to 1500 Hz (Zurek, 1993; Blauert, 1996), the octave bands centered at 500, 1000, and 2000 Hz in the STI analysis should be affected. For these three octave bands, interaural correlograms are calculated. The overall procedure is given below. Note that the choice of parameters is more fully explained in Sec. V, where we also discuss the optimization process to come to the final settings.

(a) The recorded signals (left and right ear of the artificial head) are analyzed in octave bands.
(b) For all bands, the modulation transfer function is calculated for the left and right ears separately.
(c) For the three frequency bands centered at 500, 1000, and 2000 Hz, an interaural correlogram is calculated. This is done in the following way:
    (1) The band-filtered signals are separated into nonoverlapping time frames of 30 ms duration and squared.
    (2) The left and right (squared) signals within each frame are cross correlated, resulting in a cross correlation of interaural delay for each frame (and for each filter).
    (3) Data for delay magnitudes >2 ms are discarded: the rest are kept.
    (4) Any offset in the cross-correlation function is subtracted so that the lowest value is set to zero.
    (5) Effectively, one interaural cross correlogram per frame is obtained for each band. The signal envelope can now be calculated for any interaural delay.
    (6) For a set of discrete interaural delays (τ) in the range $-2<\tau<2$, the signal power as a function of time is taken. This is already low pass filtered (with a cutoff frequency corresponding to $\frac{1}{2}$ the frame rate, i.e., 15 Hz). By using conventional techniques, the modulation transfer function (MTF) is calculated as a function of internal delay and frequency band (cf. Steeneken and Houtgast, 1980). The internal delay is selected at which the overall MTF contribution is highest (which leads to the highest STI, taking upward spread of masking into account as well). The MTF values for this internal delay are used.
(d) For the octave bands centered at 125 and 250 Hz and at 4000 and 8000 Hz, only the MTFs corresponding to left and right ears are considered. The highest value is taken (left ear or right ear) for each octave band separately.
(e) The selected (highest) MTF data from each of the seven octave bands are now combined to calculate an overall STI.

The rationale behind this approach is that for each separate band, the internal signal offering the most information, in terms of preservation of signal modulation, is presumed to be selected. How the described use of interaural correlograms would predict benefits related to interaural time delays is easily understood by considering the case of a single noise source and a single speaker, both in front of the listening position. If the speaker is slightly off to the left and the noise source to the right, then the maximum interaural correlation for the speech will be at a certain negative interaural delay, and the maximum interaural correlation for noise will be at a positive interaural delay. The power signals at these delays will have different modulation depths correspondingly.

This approach contains gross simplifications compared to accepted binaural models, such as the use of octave band filters instead of the much narrower auditory band filters. Also, instead of using inner-ear hair-cell models, we simply take the square of the signal amplitude. These choices were made in order to choose as simple a model as can be shown to work. However, the implementation of the binaural model, as described here, is only meaningful if it can be shown to yield sufficiently accurate predictions of speech intelligibility. To this end, equally balanced consonant-vowel-consonant (CVC) intelligibility tests were carried out in 39 binaural listening conditions. The validation carried out with the results from these listening tests is described below.

IV. VALIDATION OF THE BINAURAL STI

A. Speech material

The preferred method for subjective measurement of speech intelligibility in relation to the STI makes use of CVC words (Steeneken, 1992). This method uses simple nonsense words, embedded in carrier phrases, which were recorded digitally under good laboratory conditions (high quality microphones and no ambient noise). The recorded material consists of speech by eight speakers (four males and four females). Sequences of CVC test words were combined to obtain word lists of 51 words each. The source was digitally transferred to a computer, resampled to 22 kHz, and stored with 16 bit resolution. All CVC scores given in this report are the so-called *equally balanced* CVC scores. Since all phonemes have the same frequency of occurrence in the corpus of the test stimuli, the CVC score is by definition equally balanced.

The material was filtered with anechoic binaural impulse responses recorded with a Head Acoustics HMS III.2 dummy head (zero elevation and different azimuths) and with binaural impulse responses of environments (listening room, class room, and Grundtvigs cathedral) simulated in the ODEON® 7.0 software (Christensen, 2003). Speech shaped noise was also filtered with these binaural impulse responses.

B. Experimental design

The anechoic conditions all had a talker in front of the listening position and an interfering noise source (at signal-to-noise ratios of −3 and −6 dB) at positions around the head (0°, 30°, 60°, 90°, and 150°). In addition, conditions were created in which noise (correlated and uncorrelated between the ears) was directly added to the speech signals.

In the various (simulated) listening environments, realistic source and receiver positions were defined. Noise sources were also included in the simulation; Head-related transfer functions were included in the binaural room impulse responses yielded by Odeon. The overall impression

when listening to speech processed in this way was that a high degree of face value was offered by the simulations. Some conditions were included in which noise was added without binaural processing (diotic and dichotic/uncorrelated noise conditions). The speech material and noise files—a single speech source and a single noise source from various directions—were mixed electronically in different SNRs. This resulted in a set of 39 conditions. A survey of the 39 binaural conditions is given in the Appendix. The full STI signal was used for the speech source, i.e., 14 modulated signals per octave band with 7 simultaneous modulations. Noises were used as given in Table I.

The currently standardized version of the (monaural) STI has been validated by analyzing third-order polynomial fits through CVC data points (Steeneken, 1992). The same approach was now followed; however, instead of fitting a new polynomial through the data, the average monaural polynomial is plotted in each figure for comparison. Given our goal to have the binaural STI yield results that can be interpreted numerically in the same way as the existing STI versions, this seems to be a more appropriate choice.

To verify the validity of the monaural STI-CVC reference curve, a CVC test in a standard set of 40 representative monaural listening conditions was also carried out with 4 listeners, and the associated STI was calculated. Since exactly the same paradigm was used for the binaural conditions (except for the difference between diotic and binaural listening), a good correspondence between data from this monaural experiment and the reference curve serves to validate the applied implementation of the CVC test and the STI measurement. A survey of the monaural conditions is given in the Appendix.

C. Subjects

A total of seven young normal-hearing subjects (five males and two females, age range of 19–23 years) participated in the listening tests. All seven participated in the test with binaurally processed CVCs; four subjects participated in the test with monaurally processed CVCs. They were paid for their services.

D. Procedure

The processed lists were balanced for conditions and speakers and presented over headphones in a paced open-response test to the listeners, who were asked to respond by typing the perceived syllable on a computer keyboard. The individual responses were manually checked for typographic errors and inconsistencies, and then automatically processed. Hence, each data point consists of 56 speaker-listener pairs (8 speakers $\times$ 7 listeners). After the processing of the individual results, the mean equally balanced CVC score was calculated for each condition.

E. Results and discussion

The results show that the experiment covers almost the entire range of possible CVC scores, from 10% to 90% correct. The data are nicely spread between the minimum and maximum values. Results relating CVC scores and STI of the monaural conditions are given in Fig. 2. All STI values reported in this paper are obtained through full STI measurements (7 octaves and 14 modulation frequencies for each octave band).

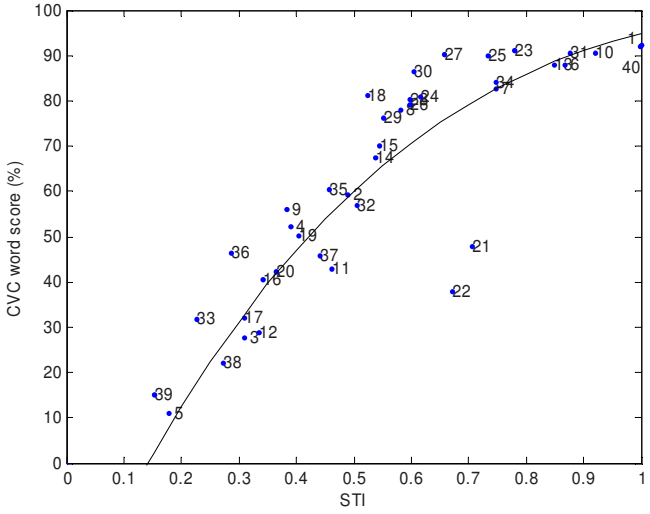

FIG. 2. (Color online) Validation of the STI vs CVC reference curve, using a set of 40 monaural reference conditions featuring noise and bandwidth limiting (1–14), nonlinear distortion (15–22), echoes (23–30), and reverberation (31–39). Condition 40 is a condition without any type of signal degradation.

Figure 2 shows that except for conditions 21 and 22, the relation between CVC and STI in monaural conditions is—on the whole—adequately described by the reference curve. Conditions 21 and 22 are center clipping conditions, for which the STI is known to overestimate intelligibility. Center clipping is nowadays rarely found in practice; it occurs with old-fashioned carbon microphones and poorly aligned push-pull amplifiers. Another noticeable deviation from the reference curve is seen at CVC scores above STI $=0.55$, where the scores appear to be approaching the saturation level more quickly. The standard deviation (or rather rms deviation), representing the vertical spread around the reference curve (cf. Steeneken, 1992), is 11.37%. This is similar to the original data set on which the reference curve is based.[1]

The relation between CVC scores and the binaural STI, in the binaural conditions described above, is shown in the top panel of Fig. 3. For comparison, the mean-ear STI in these conditions, averaged between both ears of the artificial head, is given in the middle panel, and the better-ear STI (i.e., the highest STI value of either left or right ear, as processed across all octave bands) in the bottom panel. Figure 3 shows that the relation between the binaural STI and the CVC word score comes quite close to the reference curve; the standard deviation is 9.2%, which is even smaller than for the monaural conditions of Fig. 2. For most conditions, the binaural STI seems to underestimate the intelligibility somewhat, with the exception of a cluster of data points in the cathedral environment (18–21) for which the STI is overestimated. The mean-ear STI (middle panel of Fig. 3) clearly underestimates the intelligibility in these binaural conditions with a standard deviation of 28.3%. The better-ear STI as in Fig. 3 also underestimates the binaural intelligibility. The

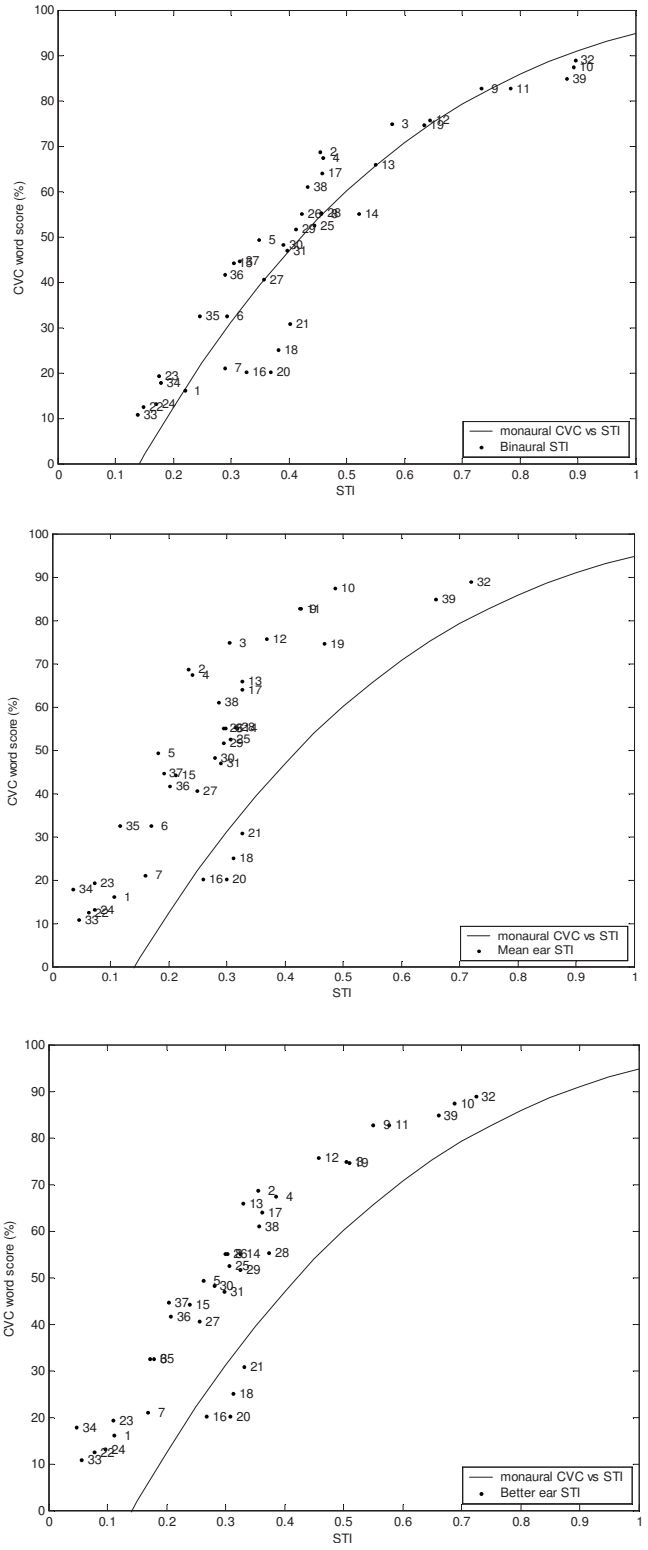

FIG. 3. CVC word score (seven subjects) as a function of the binaural STI (top panel), mean-ear STI, averaged between both ears of the artificial head (middle panel), and better-ear STI (bottom panel). Binaural conditions include anechoic conditions (1–14), a cathedral environment (15–21), a classroom (22–32), and a listening room (33–39).

standard deviation is 21.2%, which is considerably worse than for the binaural model. The better-ear STI does not take the ITD effect into account. In general, we can estimate the ITD effect to a maximum of 3 dB (difference between diotic/correlated and dichotic/uncorrelated interference). By taking the better-ear STI with a 3 dB "correction"—corresponding to a horizontal right-hand shift of 0.1 STI in Fig. 3—the data points come closer to the monaural reference curve. As a consequence, the standard deviation decreases significantly to 10.6%, quite close to the 9.2% of our binaural model.

To investigate the data further, the relation between binaural STI and CVC is given in Fig. 4 for separate categories of conditions. In various environments (anechoic, classroom, and listening room), conditions were included for which the noise presented diotically was either identical (maximum correlation) or uncorrelated. These data points are presented in a separate curve in Fig. 4.

In the anechoic conditions, the binaural STI is slightly underestimated at lower speech-to-noise ratios (−6 dB). In the cathedral environment, the binaural STI performs poorly in some cases, as observed before. This turns out to be at very large source-receiver distances (>30 m). Fortunately, such conditions are quite rare in real life. Here, more accurate estimates are actually obtained by taking the better-ear STI.

For the correlated/identical noise conditions, one would expect a difference between the anechoic conditions and to the conditions in simulated acoustic environments. Identical noise at both ears creates a clear "peak" in the interaural correlogram around an internal delay of 0 ms; uncorrelated noise contributes more or less equally at all internal delays. In an anechoic environment, speech originated from a source azimuth of 0°, straight in front of the listener position. Hence, the interaural correlation is optimal for an internal delay of 0 ms and diotic noise is expected to be a more effective masker than uncorrelated noise. However, in reverberant environments, speech signal contributions are spread out across a range of internal delays. In this case, diotic noise is expected to be less effective since the listener can "listen around" the noise peak at 0 ms (in terms of our model, the maximum STI is realized at internal delays other than 0 ms). In summary, when we subtract the intelligibility for uncorrelated noise from that for diotic noise, a positive value is expected in reverberant environments and a negative value in the anechoic environment. This is also the result found in the CVC experiment: The binaural STI correctly predicts a positive difference in reverberant conditions (+4.5% CVC difference for +0.034 STI difference) and a negative difference in the anechoic conditions (−11% CVC difference for −0.021 STI difference). However, the magnitudes of the differences are not predicted very well.

V. GENERAL DISCUSSION

Overall, the results presented in Figs. 3 and 4 appear satisfactory. However, an important question is to which degree the results presented here are influenced by the choice for the model parameters. In the proposed version, the binaural STI model has only a few free parameters, which are (a) octave bands to include in the binaural interaction model (500, 1000, and 2000 Hz), (b) range of internal delays to consider (−2–2 ms), (c) operator used to selected MTF contributions (maximum STI contribution), and (d) frame rate.

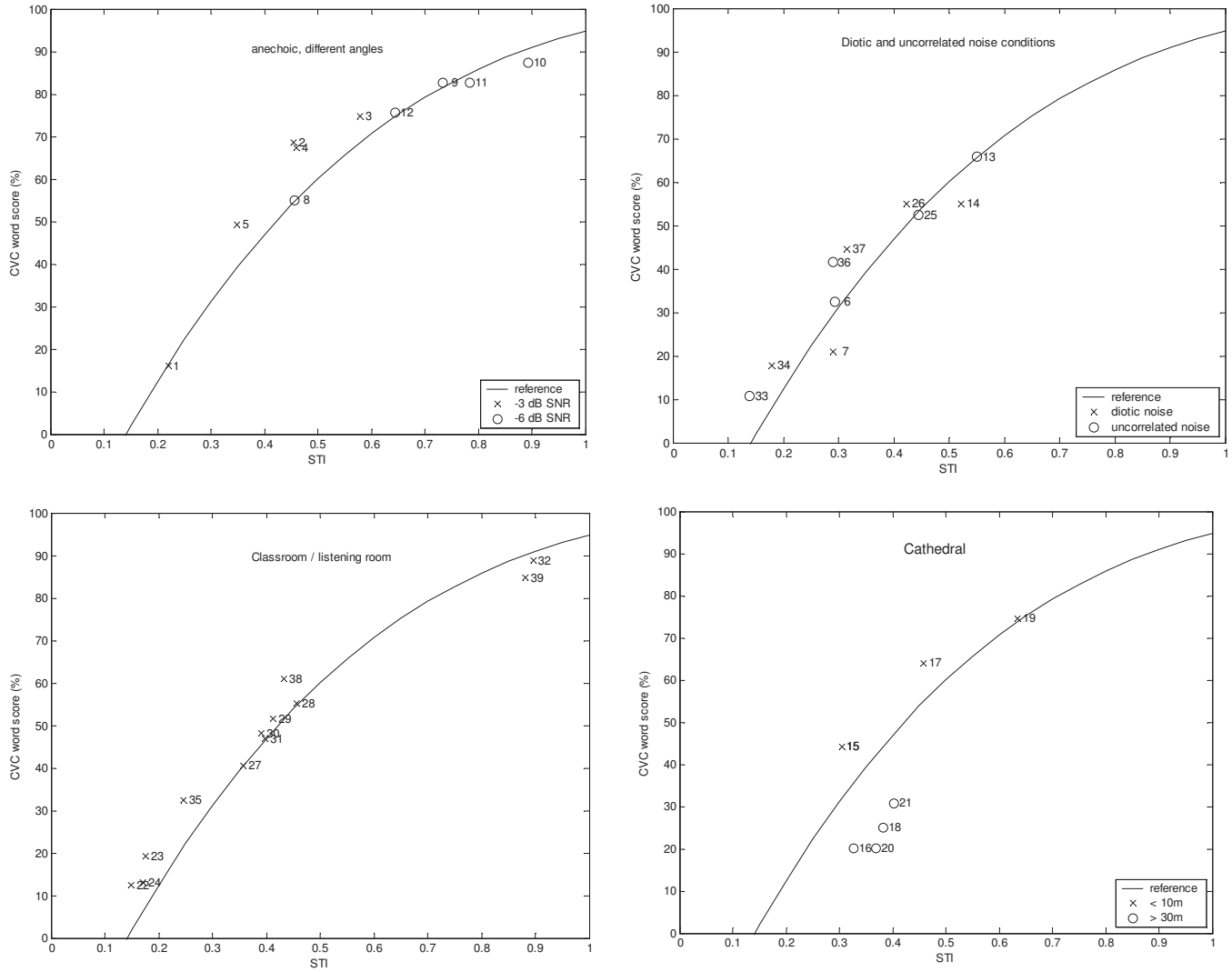

FIG. 4. Relation between CVC and binaural STI, shown separately for subsets of the binaural listening conditions (see Table I for the ID per condition). The anechoic conditions varied with respect to the noise source position. The conditions labeled "diotic noise" and "uncorrelated noise" represent various acoustic environments but with a noise signal added to both ears that is either the same or uncorrelated (no convolution with binaural impulse responses). The classroom and listening room conditions represent various source and listener positions. The cathedral conditions differed mainly with respect to the distance between the source and the receiver (here grouped in two distinct categories).

A. Octave bands

The choice which octave bands are included in the binaural interaction analysis follows from known limits of the binaural system reported in the literature. In particular, the choice to include the 2 kHz octave may be considered questionable since binaural interaction is normally presumed relatively ineffective at these frequencies, although certainly present in the lower half of the octave band. Introduction of a frequency weighting mechanism, which could be used to solve this dilemma, will only be considered as a last resort since it adds free (tunable) parameters to the model. Figure 5 shows that leaving out the 2 kHz band only slightly affects the results. Leaving out the 2 kHz *and* the 500 Hz bands clearly leads to less accurate results.

B. Internal delays

To investigate the effect of the range of internal delays taken into account, STI calculations were performed for various choices of this range. Theory predicts that interaural de-

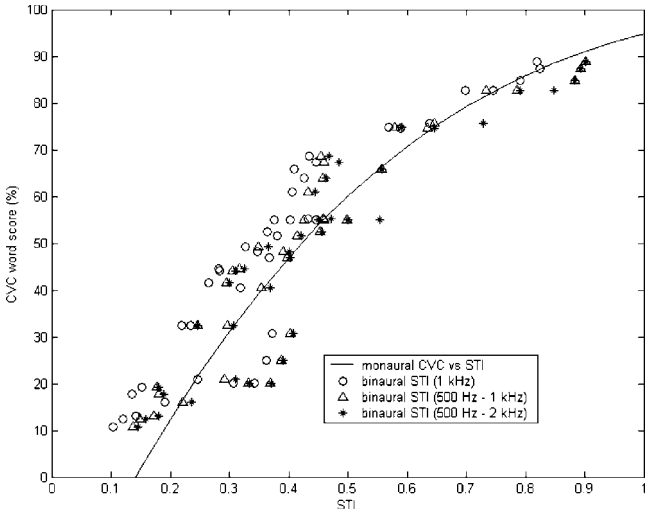

FIG. 5. CVC word score (seven subjects) as a function of the binaural STI for which binaural interaction is taken into account for one (1 kHz), two (500 Hz–1 kHz), or three (500 Hz–2 kHz) octave bands.

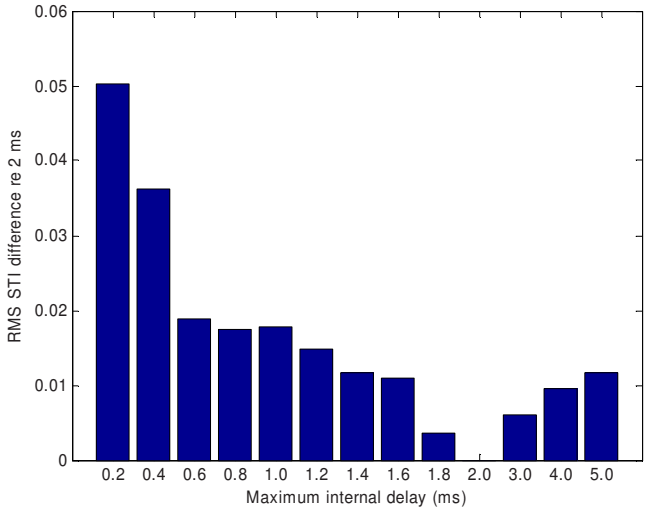

FIG. 6. (Color online) Mean absolute difference between binaural STI with a maximum internal delay of 2 ms (default) and other maximum internal delay settings. The mean is taken across the entire set of binaural conditions.

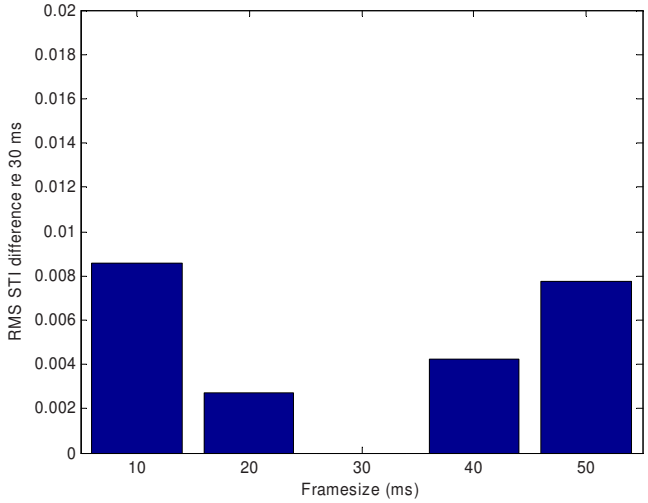

FIG. 7. (Color online) Mean absolute difference between binaural STI with frame size of 30 ms and other frame sizes. The mean is taken across the entire set of binaural conditions.

lays greater than 1 ms cannot be used effectively to enhance our internal representation of the signal (e.g., Raatgever and Bilsen, 1986). Maximum interaural delays occur for sound sources that are on the horizontal plane at an azimuth of 90° or −90°. The approximate interaural time difference is then calculated by the distance between the ears and the speed of sound; this is approximately 0.5 ms. Assuming that intelligibility benefits due to binaural interaction are limited to ecologically feasible interaural time differences, including internal delays greater than, say, 1 ms, should not result in an increased binaural STI. Figure 6 shows that this is exactly how the model behaves. Across the entire set of binaural conditions, the mean absolute difference was calculated between the binaural STI with our default internal delay range (±2 ms) and the binaural STI at various other internal delay ranges (±0.1–5 ms).

Keeping in mind that differences up to 0.03 STI occur "naturally" in monaural STI measurements due to the normal measurement error, Fig. 6 shows that it does not make a great difference whether internal delays are taken into account up to 1, 2, or 3 ms. However, if the range of internal delays is limited to a smaller maximum than, say, 1 ms, the calculated binaural STI becomes somewhat less accurate. The standard deviation relative to the monaural reference curve is, as stated above, 9.2% for the default internal delay setting (2 ms). For 0.4 ms, this standard deviation increases to 11.1%, for 0.2 ms to 12.1%, and for 0.1 ms to 21.0%.

C. MTF contributions

The choice to take the maximum of the MTF for any internal delay (instead of, for instance, the mean or median) results from the hypothesis that our binaural system selectively tunes into areas of the binaural correlogram where most information is available.

D. Frame rate

The frame rate is a parameter that may appear to be freely adjustable but for which the choices are limited for computational reasons. The STI method measures modulation frequencies up to 12.5 Hz. This means that the signal envelopes extracted from the sequence of binaural correlograms must be reliable up to this frequency, imposing a minimum frame rate of 25 Hz. Thus, the frame size must be less than 40 ms. On the other hand, the frame size must not be too small to prevent loss of accuracy in determining the interaural delays. The STI method uses a decimating filter bank to filter the signal into octave bands. This means that the signal in the 500 Hz band is sampled at 2756 Hz if the original sampling frequency is 44 100 Hz. When working with a frame size of 30 ms (our default), then this comes down to 83 samples per frame, which turns out to produce cross-correlation functions of acceptable accuracy. Shorter frames will lead to less accurate estimates of the cross-correlation function.

To determine the effect of frame size on the binaural STI, calculations were carried out similar to Fig. 3 but at frame sizes of 10, 20, 40, and 50 ms instead of 30 ms.

The effect on the calculated STI appears to be small. STI values computed on the basis of 10, 20, 40, or 50 ms are virtually identical to the values computed with a 30 ms frame size. To show this more clearly, the mean absolute difference was calculated (across the entire set of binaural conditions) between the binaural STI calculated with various frame sizes and the default frame size of 30 ms. Results are shown in Fig. 7. The mean absolute difference between monaural across STI measurements in the same condition is normally, due to measurement error alone, around 0.03. In this light, the effect of frame size is relatively minor. So, it seems fair to conclude that the model is not overly sensitive to the choice of the frame size (or the corresponding frame rate).

VI. CONCLUSIONS

When using the standard speech transmission index to predict speech intelligibility in binaural listening conditions, the intelligibility is underestimated. Significant improvement is already obtained by simply doing a two-channel STI mea-

TABLE I. Survey of the 39 binaural conditions in different environments, speech and noise positions, and signal-to-noise ratios.

Condition	ID	Speech azimuth at distance	Noise azimuth at distance	SNR (dB)
Anechoic	1, 8	0° at 1.1 m	0° at 1.1 m	−6, −3
	2, 9	0° at 1.1 m	150° at 1.1 m	−6, −3
	3, 10	0° at 1.1 m	30° at 1.1 m	−6, −3
	4, 11	0° at 1.1 m	60° at 1.1 m	−6, −3
	5, 12	0° at 1.1 m	90° at 1.1 m	−6, −3
	6, 13	0° at 1.1 m	Dichotic	−6, −3
	7, 14	0° at 1.1 m	Diotic	−6, −3
Listening room (T30≈0.4 s)	33, 36	0° at 2.6 m	Dichotic	−6, −3
	34, 37	0° at 2.6 m	Diotic	−6, −3
	35, 38	0° at 2.6 m	90° at 0.8 m	−6, −3
	39	0° at 2.6 m	No noise	∞
Classroom (T30≈0.5–1 s)	22	300° at 4.4 m	302° at 2.9 m	−6
	23	300° at 4.4 m	0° at 7.5 m	−6
	24	0° at 7.5 m	300° at 4.4 m	−6
	25, 26	0° at 2.0 m	Dichotic, diotic	−3
	27	0° at 2.0 m	0° at 2.0 m	−3
	28	0° at 2.0 m	320° at 4.2 m	−3
	29	0° at 2.0 m	230° at 3.4 m	−3
	30	0° at 2.0 m	180° at 2.2 m	−3
	31	0° at 2.0 m	140° at 2.9 m	−3
	32	0° at 2.0 m	No noise	∞
Cathedral (T30≈1.5–14 s)	15, 17	260° at 7 m	355° at 38 m	0, +3
	16, 18	345° at 38 m	270° at 7 m	0, +3
	19	345° at 38 m	No noise	∞
	20	5° at 31 m	No noise	∞
	21	330° at 33 m	No noise	∞

TABLE II. Survey of the 40 monaural conditions with different bandpass, nonlinear (peak and center clipping), and echo conditions in various signal-to-noise ratios. The peak clip level is −24 dB below the 1% speech peak level. Center clipping conditions 21 and 22 have clip levels −24 and −21 dB below the 1% speech peak level, respectively.

Condition	ID	Bandwidth (Hz)	Noise type	SNR (dB)	Echo (ms)	RT60 (ms)
Unprocessed	40	10–16 000	⋯	∞	⋯	⋯
Bandpass only	1	50–10 500	⋯	∞	⋯	⋯
	2, 3	50–10 500	White	0, −8	⋯	⋯
	4, 5	50–10 500	Pink	0, −8	⋯	⋯
	6, 7	50–10 500	Low	3, −3	⋯	⋯
	8, 9	50–10 500	Speech	3, −3	⋯	⋯
	10	300–3 400	⋯	∞	⋯	⋯
	11	300–3 400	White	0	⋯	⋯
	12	300–3 400	Pink	0	⋯	⋯
	13	300–3 400	Low	3	⋯	⋯
	14	300–3 400	Speech	3	⋯	⋯
Peak clip (+bandpass)	15	50–10 500	⋯	∞	⋯	⋯
	16	50–10 500	White	6	⋯	⋯
	17	50–10 500	Speech	3	⋯	⋯
	18	300–3 400	⋯	∞	⋯	⋯
	19	300–3 400	White	6	⋯	⋯
	20	300–3 400	Speech	6	⋯	⋯
Center clip	21,22	50–10 500	⋯	∞	⋯	⋯
Echo (+bandpass)	23	50–10 500	⋯	∞	50	⋯
	24	50–10 500	Speech	6	50	⋯
	25	50–10 500	⋯	∞	100	⋯
	26	50–10 500	Speech	6	100	⋯
	27	50–10 500	⋯	∞	200	⋯
	28	50–10 500	Speech	12	200	⋯
	29	50–10 500	Speech	6	200	⋯
	30	300–3 400	⋯	∞	200	⋯
Reverberation	31	50–10 500	⋯	∞	⋯	200
	32,33	50–10 500	Speech	6, −3	⋯	200
	34	50–10 500	⋯	∞	⋯	500
	35,36	50–10 500	Speech	6, 0	⋯	500
	37	50–10 500	⋯	∞	⋯	2000
	38,39	50–10 500	Speech	6, 0	⋯	2000

surement using an artificial head and working with the better-ear STI. However, in some conditions, this simple approach still considerably underestimates the actual intelligibility.

On the basis of the 39 binaural conditions tested in this paper, the proposed binaural STI model is capable of predicting binaural speech intelligibility with the same approximate accuracy offered by the traditional STI in monaural listening conditions. We also found that, overall, a simple better-ear STI appears to perform quite well in relation to the binaural STI model. The attractiveness of this particular binaural STI lies in a few features.

(a) The model is motivated by the existing binaural theory, considerably simplified.
(b) There are only a few "free" model parameters (frequency range and internal delay range).
(c) Changing these model parameters within reasonable bounds has little effect on the outcome of the model.
(d) The model is simple and computationally inexpensive.
(e) Known subjective binaural intelligibility data are accurately predicted by the model.

The fact that the model is relatively insensitive to changes in the model parameter values increases confidence in the strength of the model itself; it reduces that likelihood that the correspondence between subjective data and predicted intelligibility is the result of "fitting" rather than "modeling."

ACKNOWLEDGMENTS

This research was supported by grants from the European Union FP6, Project No. 004171 HEARCOM. The authors wish to thank Bastiaan van Gils for conducting the CVC experiments and Claus Lynge from Ørsted DTU, Acoustic Technology, Technical University of Denmark for providing us with the binaural impulse responses of the listening room, classroom, and cathedral.

APPENDIX: SURVEYS OF THE CONDITIONS USED IN THE EVALUATION

Tables I and II above give a survey of the binaural and

monaural signal processing conditions used in the CVC evaluation experiments described in Sec. IV.

[1]The actual standard deviations reported by Steeneken (1992) were somewhat lower (up to 8%), but these were calculated separately by category of distortions; the reference curve was fitted individually to each category. Also, the center clipping points were excluded from the standard deviation calculation. When calculated in the same straightforward way applied here, the standard deviation for Steeneken's data is about 12%.

ANSI (**1997**). "Methods for calculation of the speech intelligibility index," ANSI Report No. S3.5-1997, American National Standards Institute, New York.

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**, 331–342.

Blauert, J., (**1996**). *Spatial Hearing* (MIT, Cambridge, MA), Chap. 4, pp. 313–324.

Bronkhorst, A. W., (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acust. Acta Acust. **86**, 117–128.

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**, 1508–1516.

Bronkhorst, A. W., and Plomp, R., (**1990**). "A clinical test for the assessment of binaural speech perception in noise," Audiology **29**, 275–285.

Cherry, E. C., (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979.

Christensen, C. L., (**2003**). *Odeon Room Acoustics Program Version 6.5 User Manual* (Technical University of Denmark, Lyngby).

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

Colburn, H. S., (**1995**). "Computational models in binaural processing," in *Auditory Computation*, edited by H. Hawkins and T. McMullin (Springer, New York).

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

Durlach, N. I., (**1972**). "Binaural signal detection: Equalization and cancellation theory," in *Foundations of Modern Auditory Theory*, edited by J. V. Tobias (Academic, New York), pp. 369–462.

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

Houtgast, T., and Steeneken, H. J. M., (**2002**). "The roots of the STI approach," in *Past, Present and Future of the Speech Transmission Index*, edited by S. J. van Wijngaarden (TNO Human Factors, Soesterberg).

Houtgast, T., Steeneken, H. J. M., and Plomp, R., (**1980**). "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics," Acustica **46**, 60–72.

IEC (**2003**). "Sound system equipment. Part 16: Objective rating of speech intelligibility by speech transmission index," IEC Standard 60268-16 (3rd edition), International Electrotechnical Commission, Geneva Switzerland.

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

Kryter, K. D., (**1962**). "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**, 1689–1697.

Raatgever, J., and Bilsen, F. A., (**1986**). "A central spectrum theory of binaural processing. Evidence from dichotic pitch," J. Acoust. Soc. Am. **80**, 429–441.

Steeneken, H. J. M., (**1992**). Ph.D. thesis, University of Amsterdam, Amsterdam.

Steeneken, H. J. M., and Houtgast, T., (**1980**). "A physical method for measuring speech transmission quality," J. Acoust. Soc. Am. **67**, 318–326.

Stern, R. M., and Trahiotis, C., (**1995**). "Models of binaural interaction," in *Hearing*, edited by B. C. J. Moore (Academic, London), pp. 347–386.

Zurek, P. M., (**1993**). "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance*, 2nd ed., edited by G. A. Studebaker and I. Hochberg (Allyn and Bacon, London), Chap. 15, pp. 255–276.

Zwicker, E., and Henning, G. B., (**1985**). "The four factors leading to binaural masking-level differences," Hear. Res. **19**, 29–47.