Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model[a)]

Tim Jürgens and Thomas Brand
Medizinische Physik, Universität Oldenburg, D-26111 Oldenburg, Germany

(Received 6 June 2008; revised 2 June 2009; accepted 16 August 2009)

This study compares the phoneme recognition performance in speech-shaped noise of a microscopic model for speech recognition with the performance of normal-hearing listeners. "Microscopic" is defined in terms of this model twofold. First, the speech recognition rate is predicted on a phoneme-by-phoneme basis. Second, microscopic modeling means that the signal waveforms to be recognized are processed by mimicking elementary parts of human's auditory processing. The model is based on an approach by Holube and Kollmeier [J. Acoust. Soc. Am. **100**, 1703–1716 (1996)] and consists of a psychoacoustically and physiologically motivated preprocessing and a simple dynamic-time-warp speech recognizer. The model is evaluated while presenting nonsense speech in a closed-set paradigm. Averaged phoneme recognition rates, specific phoneme recognition rates, and phoneme confusions are analyzed. The influence of different perceptual distance measures and of the model's *a-priori* knowledge is investigated. The results show that human performance can be predicted by this model using an optimal detector, i.e., identical speech waveforms for both training of the recognizer and testing. The best model performance is yielded by distance measures which focus mainly on small perceptual distances and neglect outliers.
 [DOI: 10.1121/1.3224721]

PACS number(s): 43.71.An, 43.66.Ba, 43.71.Es, 43.72.Dv [MSS] Pages: 2635–2648

I. INTRODUCTION

The methods usually used for speech intelligibility prediction are index-based approaches, for instance, the articulation index (AI) (ANSI, 1969), the speech transmission index (STI) (Steeneken and Houtgast, 1980), and the speech intelligibility index (SII) (ANSI, 1997). AI and SII use the long-term average frequency spectra of speech and noise separately and calculate an index that can be transformed into an intelligibility score. The parameters used for the calculation are tabulated and mainly fitted to empirical data. These indices have been found to successfully predict speech intelligibility for normal-hearing subjects within various noise conditions and in silence (e.g., Kryter, 1962; Pavlovic, 1987). The STI is also index-based and uses the modulation transfer function to predict the degradation of speech intelligibility by a transmission system. All of these approaches work "macroscopically," which means that macroscopic features of the signal like the long-term frequency *spectrum* or the *signal-to-noise ratios* (SNRs) in different frequency bands are used for the calculation. Detailed *temporal* aspects of speech processing that are assumed to play a major role within our auditory speech perception are neglected. Some recent modifications to the SII improved predictions of the intelligibility in fluctuating noise (Rhebergen and Versfeld, 2005; Rhebergen *et al.*, 2006; Meyer *et al.*, 2007b) and included aspects of temporal processing by calculating the SII based on short-term frequency spectra of speech and noise. However, even these approaches do not mimic all details of auditory preprocessing that are most likely involved in extracting the relevant speech information. Furthermore, the model approaches mentioned above are "macroscopic" in a second sense as they usually predict average recognition rates of whole sets of several words or sentences and not the recognition rates and confusions of single phonemes.

The goal of this study is to evaluate a "microscopic" speech recognition model for normal-hearing listeners. We define microscopic modeling twofold. First, the *particular stages* involved in the speech recognition of normal-hearing human listeners are modeled in a typical way of psychophysics based on a detailed "internal representation" (IR) of the speech signals. Second, the recognition rates and confusions of *single phonemes* are compared to those of human listeners. This definition is in line with Barker and Cooke (2007), for instance. In our study, this kind of modeling is aimed at understanding the factors contributing to the perception of speech in normal-hearing listeners and may be extended to other acoustical signals or to understanding the implications of hearing impairment on speech perception (for an overview see, e.g., Moore, 2003).

Toward this goal we use an auditory preprocessing based on the model of Dau *et al.* (1996a) that processes the signal waveform. This processed signal is then recognized by a dynamic-time-warp (DTW) speech recognizer (Sakoe and Chiba, 1978). This is an approach proposed by Holube and Kollmeier (1996). The novel aspect of this study compared to Holube and Kollmeier (1996) is that the influence of different perceptual distance measures used to distinguish between phonemes within the speech recognizer is investigated in terms of the resulting phoneme recognition scores. Furthermore, we evaluate the predictions of this model on a phoneme scale, which means that we compare confusion ma-

[a)]Parts of this research were presented at the eighth annual conference of the International Speech Communication Association (Interspeech 2007, Antwerp, Belgium).

trices as well as overall speech intelligibility scores. This is a method commonly used in automatic speech recognition (ASR) research.

A. Microscopic modeling of speech recognition

There are different ways to predict speech intelligibility using auditory models. Stadler *et al.* (2007) used an information-theory approach in order to evaluate preprocessed speech information. This approach predicts the speech reception threshold (SRT) very well for subjects with normal hearing for a Swedish sentence test. Another way was presented by Holube and Kollmeier (1996) who used a DTW speech recognizer as a back-end to the auditory model proposed by Dau *et al.* (1996a). They were able to predict speech recognition scores of a rhyme test for listeners with normal hearing and with hearing impairment with an accuracy comparable to that of AI and STI. Both Stadler *et al.* (2007) and Holube and Kollmeier (1996) used auditory models that were originally fitted to other psychoacoustical experiments, such as masking experiments of non-speech stimuli, for instance.

Several studies indicate that temporal information is essential for speech recognition. Chi *et al.* (1999) and Elhilali *et al.* (2003), for instance, compared the predictions of a spectro-temporal modulation index to the predictions of the STI and showed that spectro-temporal modulations are crucial for speech intelligibility. They concluded that information within speech is not separable into a temporal-only and a spectral-only part but that also joint spectro-temporal dimensions contribute to overall performance. Christiansen *et al.* (2006) showed that temporal modulations of speech play a crucial role in consonant identification. For these reasons, this study uses a slightly modified version of the approach by Holube and Kollmeier (1996). The modification is a modulation filter bank (Dau and Kohlrausch, 1997) extending the perception model of Dau *et al.* (1996a), which gives the input for the speech recognition stage. It provides the recognizer with information about the modulations in the different frequency bands. The whole auditory model is based on psychoacoustical and physiological findings and was successful in describing various masking experiments (Dau *et al.*, 1996b), modulation detection (Dau and Kohlrausch, 1997), speech quality prediction (Huber and Kollmeier, 2006), and aspects of timbre perception (Emiroğlu and Kollmeier, 2008). Using a speech recognizer subsequently to the auditory model, as proposed by Holube and Kollmeier (1996), allows for predicting the SRT of an entire speech test. This approach can certainly not account for syntax, semantics, and prosody that human listeners take advantage of. To rule out these factors of human listeners' speech recognition, in the experiments of this study nonsense speech material is presented in a closed response format. The use of this speech material provides a fair comparison between the performance of human listeners and the model (cf. Lippmann, 1997). Furthermore, a detailed analysis of recognition rates and confusions of single phonemes is possible. Confusion matrices can be used in order to compare phoneme recognition rates and phoneme confusions between both humans and model results. Confusion matrices, like those used by Miller and Nicely (1955), can also be used to compare recognition rates between different phonemes provided that systematically composed speech material such as logatoms (short sequences of phonemes, e.g., vowel-consonant-vowel-utterances) is used.

The nonsense speech material of the Oldenburg logatom (OLLO) corpus (Wesker *et al.*, 2005), systematically composed from German vowels and consonants, is used for this task. This corpus was used in a former study (cf. Meyer *et al.*, 2007a) to compare human's speech performance with an automatic speech recognizer. The OLLO speech material in the study of Meyer *et al.* (2007a) allowed excluding the effect of language models that are often used in speech recognizers. Language models store plausible possible words and can use this additional information to crucially enhance the performance of a speech recognizer. Nonsense speech material was also used, for instance, in speech and auditory research to evaluate speech recognition performance of hearing impaired persons (Dubno *et al.*, 1982; Zurek and Delhorne, 1987) and to make a detailed performance comparison between automatic and human speech recognition (HSR) (Sroka and Braida, 2005). Furthermore, nonsense speech material was used, for instance, to evaluate phonetic feature recognition (Turner *et al.*, 1995) and to evaluate consonant and vowel confusions in speech-weighted noise (Phatak and Allen, 2007).

B. *A-priori* knowledge

A model for the prediction of speech intelligibility which uses an internal ASR stage deals with the usual problems of such ASR systems: error rates are much higher than those of normal-hearing human listeners in clean speech (cf. Lippmann, 1997; Meyer and Wesker, 2006) and in noise (Sroka and Braida, 2005; Meyer *et al.*, 2007a). Speech intelligibility models without an ASR stage, e.g., the SII, are provided with more *a-priori* information about the speech signal. The SII "knows" which part of the signal is speech and which part of the signal is noise because it gets them as separate inputs, which is an unrealistic and "unfair" advantage over models using an ASR stage.

For modeling of HSR the problem of too high error rates when using a speech recognizer can be avoided using an "optimal detector" (cf. Dau *et al.*, 1996a) which is also used in many psychoacoustical modeling studies. It is assumed that the recognizing stage of the model after the auditory preprocessing has perfect *a-priori* knowledge of the target signal. Limitations of the model performance are assumed to be completely located in the preprocessing stage. This strategy can be applied to a speech recognizer using template waveforms (for the training of the ASR stage) that are identical to the waveforms of the test signals except for a noise component constraining the performance. Holube and Kollmeier (1996) applied an optimal detector in form of a DTW speech recognizer as a part of their speech intelligibility model using identical speech recordings that were added with different noise passages for the model training stage and for recognizing. Hant and Alwan (2003) and Messing *et al.*

(2008) also used this "frozen speech" approach to model the discrimination of speech-like stimuli. Assuming perfect *a-priori* knowledge using an optimal detector (i.e., using identical recordings as templates and as test items) is one special case of modeling human's speech perception. Another case is using different waveforms for testing and training, thus assuming only limited knowledge about the target signal. This case corresponds not to an optimal detector but to a limited one. The latter is the standard of ASR; the former is widely used in psychoacoustic modeling. In this study, we use both the optimal detector approach and a typical ASR approach. In this way it is possible to investigate how predictions of these two approaches differ and whether the first or the second method is more appropriate for microscopic modeling of speech recognition.

C. Measures for perceptual distances

Because the effects of higher processing stages (like word recognition or use of semantic knowledge) have been excluded in this study by the use of nonsense speech material, it is possible to focus on the sensory part of speech recognition. As a working hypothesis we assume that the central human auditory system optimally utilizes the speech information included in the IR of the speech signal. This information is used to discriminate between the presented speech signal and other possible speech signals. We assume that the auditory system somehow compares the incoming speech information to an internal vocabulary "on a perceptual scale." Therefore, the following questions are of high interest for modeling: what are the mechanisms of comparing speech sounds and what is the best distance measure, on a perceptual scale, for an optimal exploitation of the speech information?

For the perception of musical tones Plomp (1976) compared the perceived similarity of tones to their differences within an equivalent rectangular bandwidth (ERB) sound pressure level spectrogram using different distance measures. Using the absolute value metric, he found higher correlations than using the Euclidean metric. For vowel sounds, however, he found a high correlation using the Euclidean metric. Emiroğlu (2007) also found that the Euclidean distance is more appropriate than, e.g., a cross-correlation measure for comparison of musical tones. The Euclidean distance was also used by Florentine and Buus (1981) to model intensity discrimination and by Ghitza and Sondhi (1997) to derive an optimal perceptual distance between two speech signals. Although the Euclidean distance was preferred by these authors for modeling the perception of sound signals, especially of speech, it still seems to be useful in this study to analyze the differences occurring on the model's "perceptual scale." By using an optimal distance measure, deduced from the empirically found distribution of these differences, the model recognition performance can possibly be optimized.

II. METHOD

A. Model structure

1. The perception model

Figure 1 shows the processing stages of the perception model. The upper part of this sketch represents the training

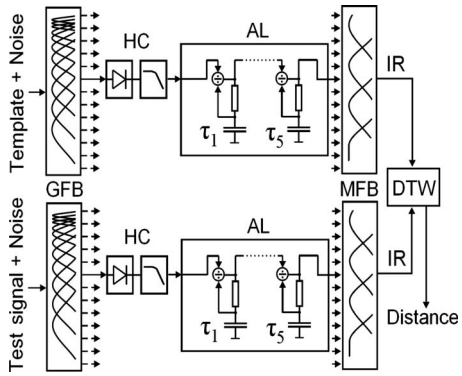

FIG. 1. Scheme of the perception model. The time signals of the template recording added with running noise and the time signal of the test signal added with running noise are preprocessed in the same effective "auditory-like" way. A gammatone filterbank (GFB), a haircell (HC) model, adaptation loops (ALs), and a modulation filterbank (MFB) are used. The outputs of the modulation filterbank are the internal representations (IRs) of the signals. They serve as inputs to the Dynamic-Time-Warp (DTW) speech recognizer that computes the "perceptual" distance between the IRs of the test logatom and the templates.

procedure. A template speech signal with optionally added background noise serves as input to the preprocessing stage. The preprocessing consists of a gammatone-filterbank (Hohmann, 2002) to model the peripheral filtering in the cochlea. 27 gammatone filters are equally spaced on an ERB-scale with one filter per ERB covering a range of center frequencies from 236 Hz to 8 kHz. In contrast to Holube and Kollmeier (1996), gammatone filters with center frequencies from 100 to 236 Hz are omitted because these filters are assumed not to contain information that is necessary to discriminate different phonemes. This is consistent with the frequency channel weighting within the calculation of the SII (ANSI, 1997) and our own preliminary results. A hearing threshold simulating noise that is spectrally shaped to human listeners' audiogram data (according to IEC 60645-1) is added to the signal before it enters the gammatone-filterbank (GFB) (cf. Beutelmann and Brand, 2006). The noise is assumed to be 4 dB above human listeners' hearing threshold for all frequencies, as proposed by Breebaart *et al.* (2001).[1] Each filter output is half-wave rectified and filtered using a first order low pass filter with a cut-off frequency of 1 kHz mimicking a very simple hair cell (HC) model. The output of this HC model is then compressed using five consecutive adaptation loops (ALs) with time constants as given in Dau *et al.* (1996a) (τ_1=5 ms, τ_2=50 ms, τ_3=129 ms, τ_4=253 ms, and τ_5=500 ms). These ALs compress stationary time signals approximately logarithmically and emphasize on- and offsets of non-stationary signals. Furthermore, a modulation filterbank (MFB) according to Dau and Kohlrausch (1997) is used. It contains four modulation channels per frequency channel: one low pass with a cut-off frequency of 2.5 Hz and three band passes with center frequencies of 5, 10, and 16.7 Hz. The bandwidths of the band pass filters are 5 Hz for center frequencies of 5 and 10 Hz, and 8.3 Hz for the band pass with center frequency of 16.7 Hz. The output of this model is an IR that is downsampled to a sampling frequency of 100 Hz. The IR thus contains a two-dimensional feature-matrix at each 10 ms time step consisting of 27 frequency channels and four modulation frequency

channels. The elements of this matrix are given in arbitrary model units (MU). Without the MFB 1 MU corresponds to 1 dB sound pressure level (SPL).

2. The DTW speech recognizer

The IR is passed to a DTW speech recognizer (Sakoe and Chiba, 1978) to "recognize" a speech sample. This DTW can be used either as an optimal detector by using a configuration that contains perfect *a-priori* knowledge or as a limited detector by withholding this knowledge (for details about these configurations see below). The DTW searches for an optimal time-transformation between the IRs of the template and the test signal by locally stretching and compressing the time axes.

The optimal time-transformation between two IRs is computed by first creating a distance matrix D. Each element $D(i,j)$ of this matrix is given by the distance between the feature-matrices of the template's IR ($\mathrm{IR}_{\mathrm{templ}}$) at time index i and the feature-matrix of the test item's IR ($\mathrm{IR}_{\mathrm{test}}$) at time index j. Different distance measures are possible in this procedure (see below). As a next step a continuous "warp path" through this distance matrix is computed (Sakoe and Chiba, 1978). This warp path has the property that averaging the matrix elements along the warp path results in a minimal overall distance. The output of the DTW is this overall distance and thus is a distance between these IRs. From an assortment of possible templates the template with the smallest distance is chosen as the recognized one.

3. Distance measures

In a first approach the Euclidean distance

$$D_{\mathrm{Euclid}}(i,j) = \sqrt{\sum_{f_{\mathrm{mod}}} \sum_{f} \left(\mathrm{IR}_{\mathrm{templ}}(i,f,f_{\mathrm{mod}}) - \mathrm{IR}_{\mathrm{test}}(j,f,f_{\mathrm{mod}})\right)^2} \tag{1}$$

between the feature-vectors $\mathrm{IR}_{\mathrm{templ}}$ and $\mathrm{IR}_{\mathrm{test}}$ was used with f denoting the frequency channel and f_{mod} denoting the modulation-frequency channel of the IRs (Jürgens *et al.*, 2007). In many studies this Euclidean distance is used when comparing perceptual differences (e.g., Plomp, 1976; Holube and Kollmeier, 1996). The Euclidean distance measure implies a Gaussian distribution of the differences between template and test IR.

As an example, Fig. 2 panel 1 shows the normalized histogram of differences Δd between the IRs ($\mathrm{IR}_{\mathrm{templ}}$ and $\mathrm{IR}_{\mathrm{test}}$) of two different recordings of the logatom /ada:/:

$$\Delta d(f,f_{\mathrm{mod}},i,j) = \mathrm{IR}_{\mathrm{templ}}(i,f,f_{\mathrm{mod}}) - \mathrm{IR}_{\mathrm{test}}(j,f,f_{\mathrm{mod}}). \tag{2}$$

In this example, the logatoms were spoken by the same male German speaker and mixed with two passages of uncorrelated ICRA1-noise (Dreschler *et al.*, 2001) at 0 dB SNR. The ICRA1-noise is a steady-state noise with speech-shaped long-term spectrum. Note that Δd corresponds to all differences occurring within a distance matrix, even those that are not part of the final warp path. However, the shape of the histogram is typical of almost all speakers and all SNRs. To investigate the shape of the histogram of differences Δd between these two IRs a Gaussian probability density function (PDF)

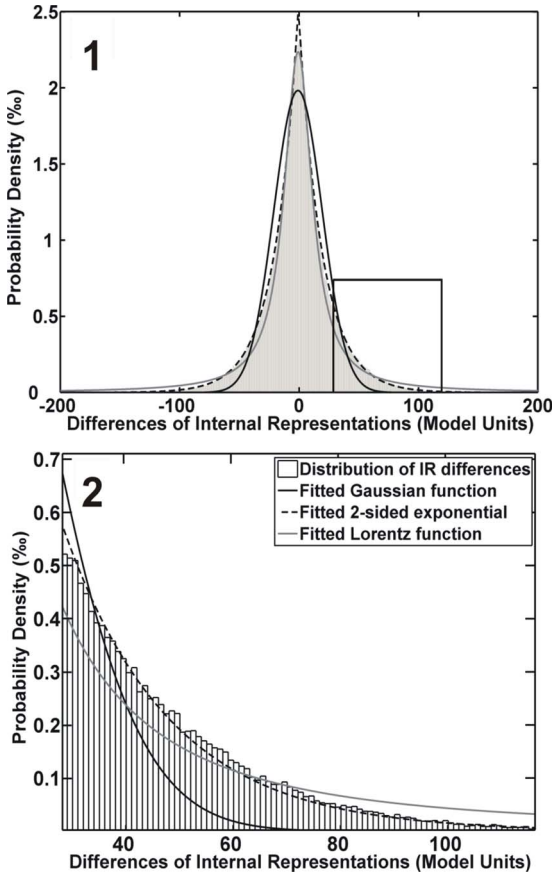

FIG. 2. (Color online) Distribution of differences (in MU) between IRs of two different recordings of the logatom /ada:/. The recordings were spoken by the same male German speaker with "normal" speech articulation style and mixed with ICRA1-noise at 0 dB SNR. A Gaussian, a two-sided-exponential, and a Lorentz-function were fitted to the data, respectively. Panel 1: complete distribution; panel 2: detail (marked rectangular) of panel 1.

$$\mathrm{PDF}_{\mathrm{Gauss}}(\Delta d) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{1}{2}\left(\frac{\Delta d_{\max} - \Delta d}{\sigma}\right)^2\right) \tag{3}$$

is fitted to the distribution which corresponds to the Euclidean metric [Eq. (1)] and a two-sided exponential PDF

$$\mathrm{PDF}_{\mathrm{exp}}(\Delta d) = \frac{1}{2\sigma} \exp\left(\left|\frac{\Delta d_{\max} - \Delta d}{\sigma}\right|\right), \tag{4}$$

and a Lorentzian PDF

$$\mathrm{PDF}_{\mathrm{Lorentz}}(\Delta d) = \frac{1}{\sqrt{2}\pi\sigma} \frac{1}{1 + \frac{1}{2}\left(\frac{\Delta d_{\max} - \Delta d}{\sigma}\right)^2} \tag{5}$$

are also fitted to the distribution, respectively. Two fitting parameters, the width of the fitted curve given by σ and the position of the maximum $\Delta d_{\max}$, must be set. The fits in Fig. 2 panel 1 show that the distribution is almost symmetrical with $\Delta d_{\max}=0$ and that high distances of about 50 MU or more are very much more frequent than expected when assuming Gaussian distributed data. Especially, very high distances of about 80 MU or more (cf. Fig. 2 panel 2) are present in the tail of outliers. The Lorentzian PDF provides a

better fit than the Gaussian function. However, it slightly overestimates the amount of outliers. The two-sided exponential function provides the best fit to the data. The two-sided exponential function is capable of reproducing the shape of the mean peak at 0 MU as well as the shape of the tail of outliers.

By taking the negative logarithm of a PDF [Eqs. (3)–(5)] and summing up the distances across all frequency channels and modulation frequency channels, a distance measure is obtained (cf. Press *et al.*, 1992) that can be used within the speech recognition process. This gives the Euclidean distance metric [Eq. (1)] (for Gaussian distributed data), the absolute value distance metric

$$D_{\mathrm{abs}}(i,j) = \sum_{f_{\mathrm{mod}}} \sum_{f} \left(\left|\mathrm{IR}_{\mathrm{templ}}(i,f,f_{\mathrm{mod}}) - \mathrm{IR}_{\mathrm{test}}(j,f,f_{\mathrm{mod}})\right|\right), \quad (6)$$

and the Lorentzian distance metric

$$D_{\mathrm{Lorentz}}(i,j) = \sum_{f_{\mathrm{mod}}} \sum_{f} \log\left[1 + \tfrac{1}{2}(\mathrm{IR}_{\mathrm{templ}}(i,f,f_{\mathrm{mod}}) - \mathrm{IR}_{\mathrm{test}}(j,f,f_{\mathrm{mod}}))^2\right]. \quad (7)$$

Note that the prefactors that normalize the PDFs are not included within Eqs. (1), (6), and (7) because they represent a constant offset in the distance metric which has no effect on the position of the minimum of the overall distance. The parameter σ is set to 1 MU for simplicity. For Eqs. (1) and (6) the value of σ is not relevant to finding the best warp path through the distance matrix (i.e., solving a constrained minimizing problem). However, in Eq. (7), σ is relevant to finding the best warp path because it cannot be factored out as it can for the Euclidean and the absolute value metric. Choosing σ equal to 1 MU results in a very flat behavior of the distance metric for middle and high distances. Other values of σ in the range from 60 to 0.1 MU showed only minor influence to the performance results in preliminary experiments.

A hypothesis for the present study is that using either Eq. (6) or Eq. (7) instead of the Euclidean distance [Eq. (1)] within the DTW speech recognition process may better account for the characteristic differences of the IRs and may improve matching.

B. Speech corpus

Speech material taken from the OLLO speech corpus (Wesker *et al.*, 2005) is used in this study. The corpus consists of 70 different vowel-consonant-vowel (VCV) and 80 consonant-vowel-consonant (CVC) logatoms composed of German phonemes. The first and the last phoneme of one logatom are the same. The middle phonemes of the logatoms are either vowels or consonants which are listed below (represented with the International Phonetic Alphabet, IPA, 1999).

- Consonants:
  /p/, /t/, /k/, /b/, /d/, /g/, /s/, /f/, /v/, /n/, /m/, /ʃ/, /ts/, /l/
- Vowels:
  /a/, /a:/, /ɛ/, /e/, /ɪ/, /i/, /ɔ/, /o/, /ʊ/, /u/

Consonants are embedded in the vowels /a/, /ɛ/, /ɪ/, /ɔ/, and /ʊ/, respectively, and vowel phonemes are embedded in the consonants /b/, /d/, /f/, /g/, /k/, /p/, /s/, and /t/, respectively.

Most of these logatoms are nonsense in German.[2] The logatoms are spoken by 40 different speakers from four different dialect regions in Germany and by ten speakers from France. The speech material covers several speech variabilities such as speaking rate, speaking effort, different German dialects, accent, and speaking style (statement and question). In the present study, only speech material of one male German speaker with no dialect and with "normal" speech articulation style is used.

C. Test conditions

Calculations with the perception model as well as measurements with human listeners were performed under highly similar conditions.

The same recordings from the logatom corpus were used. The logatoms were arranged into groups in which only the middle phoneme varied. With this group of alternatives a closed testing procedure was performed. This means that both the model and the subject had to choose from identical groups of logatoms. This allowed for a fair comparison of human and modeled speech intelligibility because the humans' semantic and linguistic knowledge had no appreciable influence. Furthermore, it allowed the recognition rates and confusions of phonemes to be analyzed. The speech waveforms were set to 60 dB SPL. Stationary noise with speech-like long-term spectrum (ICRA1-noise, Dreschler *et al.*, 2001) downsampled to a sampling frequency of 16 kHz was added to the recordings and 400 ms prior to the recording. The whole signal was faded in and out using 100 ms Hanning-ramps. After computing the IR of the speech signals as described in Secs. II A and II C, the part of it corresponding to the 400 ms noise prior to the speech signal was deleted. This was done in order to give only the information required for discriminating phonemes to the speech recognizer and not the preceding IR of the preceding background noise.

D. Modeling of *a-priori* knowledge

Two configurations of *a-priori* knowledge of the speech recognizer were realized.

- In configuration A five IRs per logatom calculated from five different waveforms were used as templates. The waveforms were randomly chosen from the recordings of one single male speaker with normal speech articulation style. None of the five waveforms underlying these IRs (the vocabulary) was identical to the tested waveform. The logatom yielding the minimum average distance between the IR of the test sample and all five IRs of the templates was chosen as the recognized one. This limited detector approach mimics a realistic task of automatic speech recognizers because the exact acoustic waveform to be recognized was unknown.
- Model configuration B used a single IR per logatom as template. The waveform of the correct response alternative

was identical to the waveform of the test signal. Thus, the resulting IRs of test signal and the correct response alternative differed only in the added background noise and hearing threshold simulating noise that were uncorrelated in time. In contrast to configuration A, this configuration disregards the natural variability of speech. Thus, it assumes perfect knowledge of the speech template to be matched using the DTW algorithm and corresponds to an optimal detector approach.

The calculation was performed ten times using different passages of background noise and hearing threshold simulating noise according to the individual audiograms of listeners participating in the experiments. The whole calculation took 100 h for configuration A (ten times for 150 logatoms at nine SNR values) and 13 h for configuration B on an up to date standard PC.

E. Subjects

Ten listeners with normal hearing (seven male, three female) aged between 19 and 37 years were employed. Their absolute hearing threshold for pure tones in standard audiometry did not exceed 10 dB hearing level (HL) between 250 Hz and 8 kHz. Only one threshold hearing loss of 20 dB at one audiometric frequency was accepted.

F. Speech tests

The recognition rates of 150 different logatoms were assessed using Sennheiser HDA 200 headphones in a sound-insulated booth. The calibration was performed using a Brüel&Kjaer (B&K) measuring amplifier (Type 2610), a B&K artificial ear (Type 4153), and a B&K microphone (Type 4192). All stimuli were free-field-equalized using an FIR-filter with 801 coefficients and were presented diotically. SNRs of 0, −5, −10, −15, and −20 dB were used for the presentation to human listeners. For each SNR a different presentation order of the 150 logatoms was randomly chosen. For this purpose, the 150 recordings were split into two lists, and the order of presentation of the recordings within the two lists was shuffled. Then all ten resulting lists of all SNRs were randomly interleaved for presentation. Response alternatives for a single logatom had the same preceding and subsequent phoneme (closed test); hence, the subject had to choose either from 10 (CVC) or 14 (VCV) alternatives. The subject was asked to choose the recognized logatom from the list and was asked to guess if nothing was understood. The order of response alternatives shown to the subject was shuffled as well. Before the main measurement all subjects were trained with a list of 50 logatoms.

For characterizing the mean intelligibility scores across all logatoms the model function

$$\Psi(x) = \frac{1-g}{1+\exp(4s(\mathrm{SRT}-L))} + g \qquad (8)$$

was fitted to the mean recognition rate (combined for CVCs and VCVs) for each SNR by varying the free parameters SRT and s (slope of the psychometric function at the SRT). The SRT is the SNR at approximately 55% recognition rate

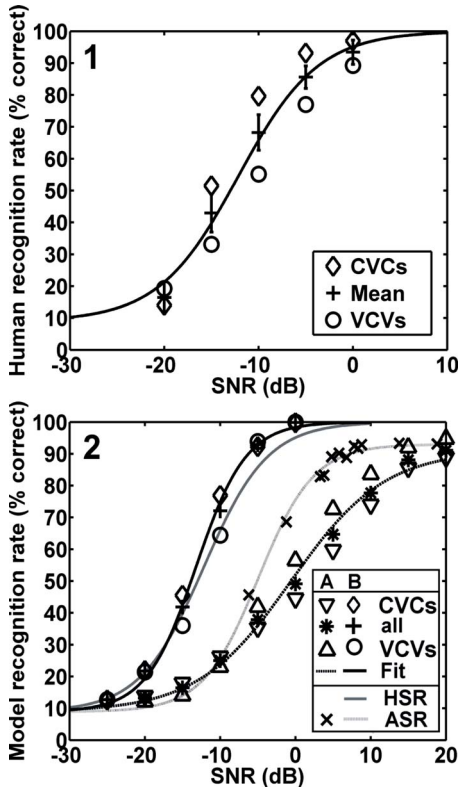

FIG. 3. (Color online) Panel 1: Psychometric function (recognition rate versus SNR) of ten normal-hearing listeners using logatoms in ICRA1-noise. Error bars correspond to the inter-individual standard deviations across subjects. Lines show the fit by Eq. (8). Panel 2: Psychometric function of the perception model with configurations A and B derived with the same utterances of the OLLO speech corpus as for the measurement. The measured psychometric function (taken from panel 1) is additionally shown for comparison as gray line (HSR). For a further comparison, data of Meyer *et al.* (2007a) are plotted (ASR).

(averaged across all CVCs and VCVs) which is the midpoint between the guessing probability and 100%. L corresponds to the given SNR and g is the guessing probability averaged across all CVCs and VCVs ($g=8.9\%$). The fit is performed by maximizing the likelihood assuming that the recognition of each logatom is a Bernoulli trial (cf. Brand and Kollmeier, 2002). Note that this fitting function assumes that 100% recognition rate is reached at high SNRs. This is feasible for listeners with normal hearing and for speech recognition modeling using an optimal detector, but is not necessarily the case for a real ASR system as such an ASR system will still show high error rates on speech material with a low redundancy even when the SNR is very high (Lippmann, 1997). For model configuration A the fitting curve is therefore fixed at the highest recognition rate that occurred in the ASR test.

III. RESULTS AND DISCUSSION

A. Average recognition rates

Figure 3 panel 1 shows the mean phoneme recognition rates in percent correct versus SNR across all phonemes. Error bars denote the inter-individual standard deviations of the ten normal-hearing subjects. Furthermore, the recognition rates of CVCs and VCVs are plotted separately. The recognition rates for CVCs are higher than for VCVs except for −20 dB SNR. The fitting of the psychometric function to

TABLE I. List of fitted parameters characterizing observed and predicted psychometric functions for the discrimination of logatoms in ICRAI noise. Rows denote different distance measures used by the DTW speech recognizer and different model configurations (see Secs. II A and II C for details) as well as values of human listeners. Pearson's rank correlation coefficients (last column) were calculated using the observed data of individual human listeners. * denotes significant ($p<0.05$) and ** highly significant ($p<0.01$) correlations.

	SRT (dB SNR)	Difference to observed SRT (dB)	Slope (%/dB)	Pearson's r^2
Human listeners	−12.2	0[a]	5.4	1[a]
Euclidean, Conf. A	−0.4	11.8	5.7	0.64**
Euclidean, Conf. B	−8.1	4.1	10.0	0.83**
Two-sided exp., Conf. A	−0.4	11.8	5.8	0.65**
Two-sided exp., Conf. B	−10.6	1.6	8.4	0.92**
Lorentzian, Conf. A	−0.6	11.6	3.5	0.83**
Lorentzian, Conf. B	−13.2	−1.0	6.8	0.97**

[a]By definition.

the data yields a slope of $5.4\pm0.6\%$/dB and a SRT of -12.2 ± 1.1 dB. Note that even the recognition rate at −20 dB SNR is significantly above chance and therefore included in the fitting procedure.

The observed and the predicted results calculated with different distance measures and model configurations are shown in Table I. The smallest differences from the observed SRT values are found for configuration B. Using this configuration, the slope of the predicted psychometric function is slightly overestimated. However, model configuration A, which performs a typical task of speech recognizers, shows a large gap of about 12 dB between predicted and observed SRTs, which is typical of ASR (see below). This gap is nearly independent of the type of distance measure, while the slope is slightly underestimated. The last column of Table I shows Pearson's squared rank correlation coefficient r^2 between the individual observed and predicted speech recognition scores. The Lorentzian distance measure using model configuration B shows the highest r^2 of 0.97 ($p<0.01$) whereas the two-sided exponential and the Euclidean distance measure show somewhat lower correlation coefficients and higher differences between observed and predicted SRTs. Different distance measures do not substantially affect the prediction of the SRT using model configuration A.

The predicted psychometric function of this best fitting model realization (configuration B with Lorentzian distance measure) is displayed in Fig. 3 panel 2. In addition, the fitted psychometric function of Fig. 3 panel 1 is replotted (HSR), and the predicted psychometric function of model configuration A with Lorentzian distance measure is shown. Furthermore, ASR-data of Meyer *et al.* (2007a) were included for comparison (see Sec. IV). For model configuration B the resulting SRT using the Lorentzian distance measure is −13.2 dB SNR and thus within the interval of the subjects' inter-individual standard deviation. The ranking of the recognition of vowels and consonants (i.e., that CVCs are better understood than VCVs) is predicted correctly except for −20 dB SNR. Model configuration A, which performs a typical task of speech recognizers, shows a SRT of −0.6 dB and

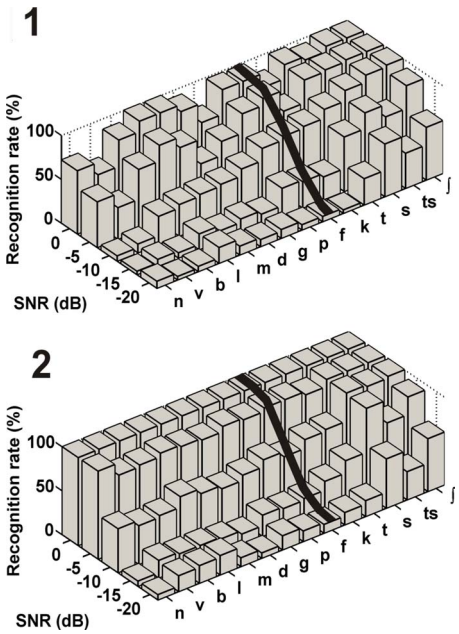

FIG. 4. (Color online) Recognition rates of consonants, separately, as a function of SNR for ten normal-hearing listeners (panel 1) and for model configuration B with Lorentzian distance measure (panel 2). As an example the psychometric function for the discrimination of /f/ in noise is shown (solid line).

a slope of 3.5%/dB using the Lorentzian distance measure. With this configuration the ranking of the recognition of vowels and consonants could not be predicted, i.e., the model shows higher recognition rates for consonants than for vowels.

B. Phoneme recognition rates at different SNRs

Figure 4 shows the recognition rates of single consonants embedded in logatoms as a function of SNR for normal-hearing listeners (panel 1) and for model configuration B using the Lorentzian distance measure (panel 2). Picking out one phoneme, the psychometric function for this specific phoneme can be seen. The solid lines in panels 1 and 2 show these psychometric functions for the phoneme /f/ as an example. Normal-hearing listeners show quite poor recognition rates for the phonemes /n/, /v/, or /g/ at the SNRs chosen for measurement. However, there are also some phonemes like /s/, /ts/, and /ʃ/ that show very high recognition rates at these SNRs. The predicted recognition rates for the latter phonemes (see panel 2) fit the observed recognition rates quite well. This is also the case for /l/, /m/, /p/, /f/, and /t/. For the other phonemes there is a discrepancy between observed and predicted recognition rates especially at high SNRs. For instance, at 0 dB SNR the predicted recognition rate is almost 100% for all phonemes, but normal-hearing listeners actually show poor recognition rates of 58% for /v/ or 70% for /g/. The recognition rates for vowels across SNR are shown in Fig. 5. Normal-hearing listeners show quite a steep psychometric function for the phonemes /e/, /ɛ/, /a:/, and /i/ but a shallower psychometric function for the other phonemes. The predicted recognition rates for /o/ and /u/ fit the observed recognition rates quite well across all SNRs investigated in this study. However, for /e/, /ɛ/, /a:/, and /i/

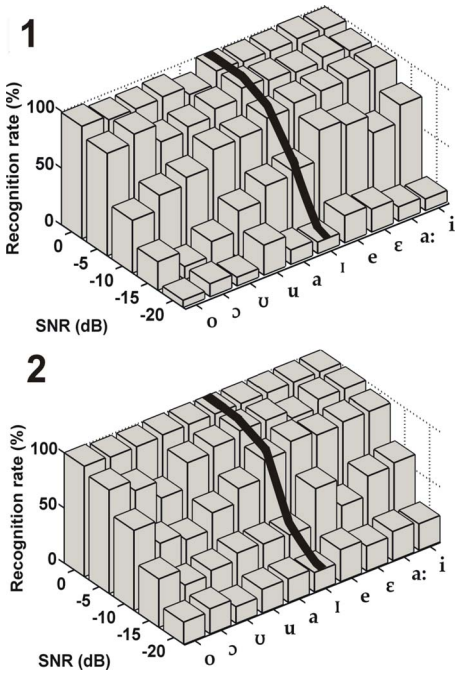

FIG. 5. (Color online) Recognition rates of vowels. The display is the same as in Fig. 4.

the predicted psychometric functions are too shallow. Note that for vowels, contrary to consonants, at 0 dB SNR almost 100% recognition rates are reached by both normal-hearing listeners and model configuration B.

C. Phoneme confusion matrices

Confusion matrices are calculated for all SNRs which were used in the experiment. In Sec. IV the confusion matrices at −15 dB SNR are analyzed. The recognition rates at this SNR are the least influenced by ceiling effects (see Figs. 4 and 5) and show the largest variation across phonemes. Therefore, at this SNR, the patterns of recognition are most characteristic. Figure 6 panel 1 shows the observed confusion matrices of the VCV discrimination task and panel 2 the corresponding predictions using the Lorentzian distance measure with model configuration B. Each row of the confusion matrix corresponds to a specific presented phoneme, and each column corresponds to a recognized phoneme. The diagonal elements denote the rates of correct recognized phonemes and the non-diagonal elements denote confusion rates of phonemes. All numbers are given in percentages.

At −15 dB SNR the average recognition rates for all consonants are 33% (human) and 36% (model configuration B, see also Fig. 3). In the following text the comparison of the two matrices will be described element-wise. Two elements differ significantly if the two-sided 95% confidence intervals surrounding the respective elements do not overlap (cf. Appendix). The observed and the predicted correct consonant recognition rates do not differ significantly, except for the phonemes /s/, /b/, and /v/. Rates below 17% do not differ significantly from the guessing probability of 7% (cf. Appendix). Hence, almost all non-diagonal elements of the model confusion matrix do not differ significantly from the corresponding elements of the human listeners' confusion matrix.

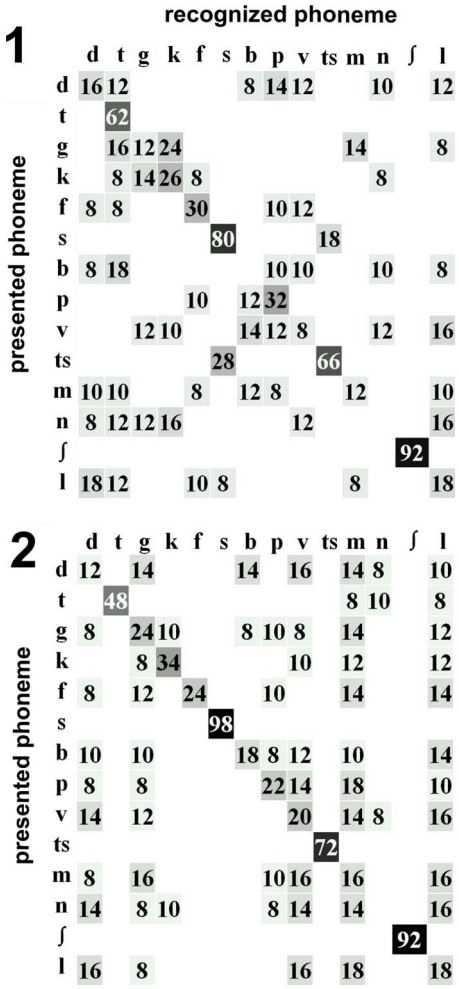

FIG. 6. (Color online) Confusion matrices (response rates in percent) for consonants at −15 dB SNR for normal-hearing subjects (panel 1) and for model configuration B with Lorentzian distance measure (panel 2). Row: presented phoneme; column: recognized phoneme. For better clarity, the values in the cells are highlighted using gray shadings with dark corresponding to high and light corresponding to low response rates. Response rates below 8% are not shown.

One exception is the confusion "presented /ts/-recognized /s/," found in the observed confusion matrix, which cannot be found in the predicted confusion matrix. Other exceptions like "presented /p/-recognized /m/" differ just significantly and shall not be discussed in detail in this section. Unfortunately, the size of confidence intervals of the matrix elements decreases very slowly with an increasing amount of data. Therefore, it is not possible to find many significant differences between predicted and observed matrix elements although the amount of data is already relatively large. However, if we compare the correct recognition rates within one matrix many phonemes can be found that differ significantly in recognition rate. Note that within one single matrix only matrix elements from different rows should be compared (cf. Appendix).

Figure 7 panel 1 shows the observed confusion matrices of the CVC discrimination task and panel B the corresponding predictions using the Lorentzian distance measure with model configuration B. At −15 dB SNR the average recognition rates for all vowels are 52% (human) and 46% (model configuration B, see also Fig. 3 panel 2). The ranking of the

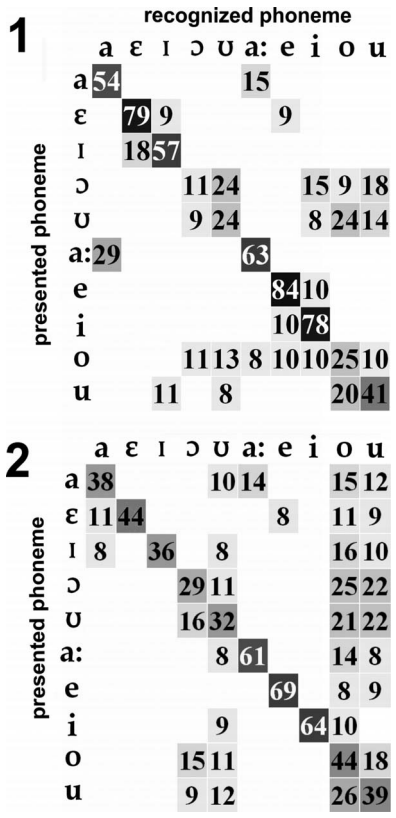

FIG. 7. Confusion matrices (response rates in percent) for vowels at −15 dB SNR for normal-hearing subjects (panel 1) and of model configuration B (panel 2). The display is the same as in Fig. 6.

best recognized phonemes /e/ and /i/, as well as the ranking of the worst recognized phonemes /o/ and /u/, is predicted correctly. However, the overall "contrast" (i.e., the difference between best and worst recognized phonemes) of the predicted matrix is much less pronounced than in the observed matrix. The largest number of confusions occurred between the phonemes /ʊ/, /ɔ/, /o/, and /u/ for both predictions and observations. However, the significant observed confusion "presented /a:/-recognized /a/" cannot be found in the predicted confusion matrix. Furthermore, the phonemes /o/ and /u/ are recognized with a bias by the model, i.e., no matter what phoneme is presented, the model shows a slight preference for these phonemes.

Pearson's φ^2 (Lancaster, 1958) index was used for comparing the similarity between measured and modeled confusion matrix data. This index is based on the chi-square test of equality for two sets of frequencies and provides a normalized measure for the dissimilarity of two sets of frequencies. A value $\varphi^2=1$ is related to complete dissimilarity whereas a value of $\varphi^2=0$ is related to equality. Table II shows φ^2 values for comparing the confusion patterns, i.e., each φ^2 value is a measure for the dissimilarity of the xth row of the observed confusion matrix and the xth row of the predicted confusion matrix of Figs. 6 and 7, respectively. For the consonant confusion matrices highest similarity is found for the confusion patterns of /t/, /s/, and /ʃ/. This very high similarity is mainly due to the high correct response, i.e., the diagonal element. Generally, many observed and predicted confusion patterns show high similarity due to low φ^2-values. However, the observed and predicted confusion patterns of /ts/ show the lowest similarity. This is mainly due to the significant confusion of "presented /ts/-recognized /s/" which was not predicted by the model. The confusion patterns of the phonemes /f/, /l/, and /p/ show moderate similarity. These phonemes also show a poor recognition rate at −15 dB SNR and thus higher percentages in the non-diagonal elements. This gives support to the supposition that the model is not able to predict the consonant *confusions* of normal-hearing listeners. For comparing the patterns of recognition, i.e., the diagonal of the confusion matrix, the correlation coefficients between observed and predicted data are shown in Table III as a function of SNR. For a SNR of −15 dB this correlation coefficient amounts to $r^2=0.91$ ($p<0.01$). This strong correlation means that the model is quite good in modeling the *correct responses*. For observed and predicted consonants there are also highly significant correlations found at −10 and −20 dB SNRs. The correlation decreases rapidly for higher SNR mainly due to ceiling effects, i.e., many phoneme recognition scores are in the range of 100%. Note that at 0 dB SNR a correlation coefficient for consonants could not be assigned due to the fact that at this SNR all consonants are predicted at a recognition rate of 100% whereas some were observed at lower recognition rates.

For the vowel confusion matrices highest similarity is found for the observed and predicted confusion patterns of /a/, /ʊ/, and /u/. Many confusion patterns show a high similarity except for those of /ɛ/, /ɔ/, and /a:/ which show only

TABLE II. Pearson's φ^2 index, a measure of dissimilarity, for comparing the confusion patterns, i.e., one row of a confusion matrix, of observed and predicted phoneme recognitions from Figs. 6 and 7, respectively.

Presented consonant	φ^2	Presented vowel	φ^2
/d/	0.21	/a/	0.10
/t/	0.12	/ɛ/	0.24
/g/	0.24	/ɪ/	0.19
/k/	0.20	/ə/	0.21
/f/	0.16	/ɔ/	0.11
/s/	0.12	/a:/	0.24
/b/	0.15	/e/	0.14
/p/	0.16	/i/	0.15
/v/	0.14	/o/	0.14
/ts/	0.25	/u/	0.10
/m/	0.21		
/n/	0.14		
/ʃ/	0.08		
/l/	0.18		

TABLE III. Correlation coefficients r^2 for comparing observed and predicted recognition scores from Figs. 4 and 5, i.e., the diagonals of confusion matrices, as a function of SNR. * denotes significant ($p<0.05$) and ** highly significant ($p<0.01$) correlations.

SNR (dB)	r^2 for consonants	r^2 for vowels
0	Not assigned	0.09
−5	0.34*	0.52*
−10	0.78**	0.56**
−15	0.91**	0.57*
−20	0.86**	0.26

modest similarity. The high similarity for the former phonemes is mainly due to the correct modeling of confusions "presented /a/-recognized /a:/", "presented /ʊ/-recognized /u/", and "presented /u/-recognized /o/", and the correct responses, respectively. The modest similarity for /ɛ/, /ɔ/, and /a:/ is mainly due to the high discrepancy in predicting the correct diagonal element score. Correlating the diagonals at this SNR (cf. also Table III) shows that the patterns of recognition are significantly ($r^2=0.57$, $p<0.05$) correlated but not as high as for the consonant recognition patterns. This also holds for −10 and −20 dB SNRs. For higher SNRs, i.e., higher average recognition scores, the correlation of predicted and observed vowels is higher than the correlation of consonants. This leads to the assumption that the model can better predict the confusion patterns for vowels than for consonants at low recognition rates as, e.g., for /ʊ/ and /u/. In predicting the correct responses, however, the model is not as good for the vowels as for the consonants.

The fact that the model is not able to predict confusion patterns correctly, especially for consonants, may be due to two reasons. The first reason may be that the model is partly not able to exploit similarities between the IRs of phonemes that might, in fact, be similar to one another for normal-hearing listeners. This is supported by a confusion that is not predicted ("presented /ts/-recognized /s/"), but not, e.g., by the confusions between /u/ and /o/ that are almost correctly predicted. The second reason may be simply due to the high ranges of confidence intervals (see Appendix) due to the inherent binomial statistics of this speech test.

IV. GENERAL DISCUSSION

A. Microscopic prediction of speech intelligibility

This study compares the recognition performance in noise of a microscopic speech intelligibility prediction model to the phoneme recognition performance of human listeners. The model was also used with the same approach as in this study to predict speech intelligibility of a rhyme test (Holube and Kollmeier, 1996). Our results, as well as the results of Holube and Kollmeier (1996), show that this combination of perception model and DTW speech recognizer is able to discriminate noisy speech signals in a closed-set testing procedure. The model used here is also similar to the microscopic model used by Barker and Cooke (2007). Their model is inspired by ASR techniques and evaluates speech parts that "glimpse" the spectro-temporal pattern of the signal to be recognized out of background noise. One main novelty of this study is that the use of the speech database of Wesker *et al.* (2005), which provides many recordings of the same logatom, allows the investigation of the influence of *a-priori* knowledge about the speech. This investigation is possible because the speech recognizer is realized with two model configurations. In model configuration B templates are used which are identical to the test items; this corresponds to maximum *a-priori* knowledge. In model configuration A the recognizer used templates which are not identical to the test items corresponding to less *a-priori* knowledge.

Assuming limited *a-priori* knowledge within model configuration A results in a much poorer performance than observed in the results of human listeners. This reflects the gap between human and machine speech reception (cf. Jürgens *et al.*, 2007) because configuration A is the standard case for ASR. The gap of about 11–12 dB SNR is consistent with findings of other studies employing common speech recognition systems like hidden-Markov-models (HMMs). Meyer *et al.* (2007a) found a gap of about 10 dB SNR (averaged across different speakers) between human listeners' SRT and the SRT of a speech recognizer using mel-frequency-cepstral-coefficients and a HMM using the same OLLO speech corpus and very similar listening experiments. As a direct comparison, a subset of the ASR-data of Meyer *et al.* (2007a) is plotted as an additional psychometric function in Fig. 3. The subset of speech material to be tested is limited to the same speech material that was used in the present study. For this speech material the gap in SRT between ASR and normal-hearing listeners' performance extends to about 8 dB. The difference of 3–4 dB from our results might be due to different speech recognizers used. Meyer *et al.* (2007a) used a speech recognizer that benefited from decades of research. Also the amount of training material in their study was much larger (49 speakers with different articulation styles) than in the present study.

Speech intelligibility can be predicted with greater accuracy using model configuration B in which the amount of information about the speech signal prior to the recognizing process is assumed to be perfect. It has to be stated that in this point the model differs from human listeners' speech processing because human listeners have not stored the exact IR of the signal to be recognized. Human listeners are able to generalize their IR of a speech utterance to different speech waveforms, even if different articulation styles or speakers are involved. However, our speech recognition model includes a pattern recognizer that has to find a speech pattern among different alternatives, which is closer to human speech processing than, for example, the SII (ANSI, 1997). This optimal detector concept is a standard in psychoacoustic modeling and predicts, e.g., forward, backward, and simultaneous masking thresholds (Dau *et al.*, 1996b), modulation detection thresholds (Dau and Kohlrausch, 1997), and the time resolution of the binaural system (Breebaart *et al.*, 2002). As this speech recognition study is in line with other psychoacoustic experiment studies because of the closed-test paradigm and the nonsense speech material used here, such an approach seems to be appropriate. The very accurate agreement of observed and predicted phoneme recognition rates using model configuration B does not mean that human listeners have a perfect decision device. Humans' limitations in discriminating speech in noise are certainly due to energetic masking of the speech signal by background noises and also due to errors in the inherent processing in the subsequent word recognition stage. However, the speech discrimination performance of the model is very similar to that of human listeners if all limitations of performance are assumed entirely in the preprocessing stage of the model. For the experiments presented here this may be interpreted as that lifelong training of humans in speech makes the pattern recognizing part of HSR perform as well as the model's optimal detector.

With configuration B the model is capable of predicting the SRT of this speech test with an accuracy of about 1 dB. The SII (ANSI, 1997) predicts the SRT within the same accuracy range: For −15 dB SNR the SII-value is found to be 0.045, for instance, and for −10 dB the SII is 0.18. Transformed to intelligibility scores by using the SII transfer function for Hagerman's sentences in noise (Magnusson, 1996), the resulting SRT is −11.2 dB SNR. The main advantage of the microscopic modeling approach compared to the SII is that, whereas the SII is able to predict only average recognition scores, this approach is able to predict the recognition scores for each phoneme separately. Furthermore, this approach draws out some characteristic phoneme confusions that are commonly seen.

B. Distance measures

The type of distance measure crucially influences the performance of the speech recognizer when using model configuration B. The Euclidean distance used by, e.g., Plomp (1976), Holube and Kollmeier (1996), and Jürgens *et al.* (2007) shows the poorest performance among the distance measures investigated here. In this study, there is a gap of more than 4 dB between the SRT of model configuration B and human listeners' SRT. Using the Euclidean distance, outlying passages are strongly weighted and consequently the DTW algorithm tries to minimize the occurrence of outlying passages as far as possible. This may cause the warp path, i.e., the temporal matching function between two IRs, to be fitted more to the passages containing different speech or noise. Passages with low distances are disregarded. By applying a distance measure that is less sensitive to outliers in the matching procedure of two IRs (i.e., using the two-sided exponential measure or the Lorentzian measure) this gap is substantially decreased or vanishes. Using the two-sided exponential distance measure, all distances are weighted with their usual occurrence probability (cf. Fig. 2). Therefore, this can be called a "natural" distance measure for speech in noise. Although no substantial influence of the type of distance measure was found on the performance of model configuration A, it was found for model configuration B. One could argue, since configuration A is typical of an ASR system, that other ASR systems may not benefit from an optimization of the distance measure they use. However, as this approach uses a speech recognizer that does not require a large amount of training material as common ASR systems do, this is speculative. Nevertheless, for further optimizing of ASR systems it may be useful to study the influence of different distance measures on the ASR systems' performance.

Using the Lorentzian distance measure, all outlying passages get approximately the same constant weight because of the flatness of the logarithm for large input values. Therefore, the overall distance between two IRs is mainly dominated by the smallest elements of the distance matrix. In other words, the steepness of the logarithm at low values causes similar passages of the IRs to be matched as closely as possible. This may particularly be an advantage for discriminating noisy speech samples because the speech recognizer is dominated by matched (i.e., similar speech) patterns and neglects unmatched (i.e., noise or different speech) patterns. Hence, the detector can separate the objects "matched speech" passages from "unmatched speech" or "noise only" passages more appropriately. If we conceive of noise and speech as different acoustical objects this mechanism may have some similarities to the mechanism of acoustical object separation within the human auditory system. Neglecting passages that do not match passages of stored response alternatives is a candidate for modeling human's mechanism of object separation. In that way the distinction between a "matchable speech object" and a "not matchable speech object" or "noise-only object" may be enhanced. Using model configuration B, the Lorentzian distance measure performs best and results in a high agreement in phoneme recognition. Therefore, this set-up was chosen for the prediction of speech recognition in noise of listeners with normal hearing.

C. Phoneme recognition rates and confusions

In this study both human listeners and the model show the highest performance at the same consonants /t/, /s/, /ʃ/, and /ts/ as in the study of Phatak and Allen (2007) who investigated consonant recognition rates in speech weighted noise. The results obtained in this study are in line with those of Phatak and Allen (2007), although they used speakers and listeners of a different native language and different speech material. Furthermore, the amount of alternatives that could be recognized was completely different from our measurements. However, the separation of consonants into a low scoring and a midscoring group with the same phonemes as in Phatak and Allen (2007) could not be observed in this study. They concluded that differences in recognition rates can mainly be explained by differences in the long-term spectra of speech and noise. However, this may not account for consonants with characteristics that are mainly determined by the temporal structure as, e.g., for plosives like /p/, /t/, or /k/. Our approach regards this temporal structure by the temporal matching performed in the DTW speech recognizer.

By and large, the confusion matrices of human listeners and of model configuration B with Lorentzian distance measure are very similar. Except for a small number of elements, the consonant confusion matrices do not differ significantly element-wise regarding the binomial statistics valid for these discrimination tasks (see Appendix). The correlation between predicted and observed recognition rates of single phonemes is very high. This is promising and it may indicate that for all phonemes speech information is conserved or emphasized during the modeled "effective" auditory preprocessing in a way similar to human listeners.

The vowel confusion matrix of the model shows a slight preference, i.e., a bias, concerning the vowels /ʊ/, /ɔ/, /o/, and /u/ independent of the presented vowel. This is one main difference between the predicted and observed vowel confusion matrices. Meyer *et al.* (2007a) found that the phonemes /o/ and /u/ within this speech corpus have the least distinctive average spectrum compared to speech-shaped noise. Consequently these phonemes are the phonemes best masked in the background noise at low SNRs. If the speech recognizer is

not able to match a presented phoneme, it is very probable that it matches the IR that is the most similar to the IR of the background noise. These are the IR of logatoms with /o/ and /u/ as middle phonemes. In some cases the procedure probably matches mainly the background noise characteristics of the IR and is not able to focus on the speech characteristics anymore. One reason why the prediction of vowel recognition rates is poorer than for consonants while the prediction of vowel confusions is better than for the consonants may be the spectro-temporal structure of these two phoneme groups. Generally, vowels are more stationary signals than consonants. Furthermore, there is no clear separation between different vowels but a continuous transition in the frequency range. Therefore, it seems reasonable to assume that two different vowels are "perceptually" closer to one another than are two different consonants. This may explain why confusions occur more frequently in both normal-hearing listeners' and modeled data.

D. Variability in the data

Data obtained by speech tests using human listeners always show both intra-individual and inter-individual variabilities. One factor for the inter-individual variability is the variability of the hearing threshold across listeners. Preliminary simulations, however, showed that adapting only the hearing threshold simulating noise results in less variability than found in normal-hearing listeners' speech recognition data. This can be explained by the low rms level of the hearing threshold simulating noise which is masked by the much higher level of the background noise. For this reason a much more effective way to include variability was to use running background noise. In other words the variability in the simulations originates almost exclusively from the statistics of the background noise. However, this is somewhat unrealistic because in the measurements the background noise stimuli were identical for every participant whereas, in reality the auditory processing varied. It still remains an open question how to obtain a comparable variability by modifying the auditory processing without using this workaround. For speech intelligibility modeling in silence, e.g., Holube and Kollmeier (1996) achieved some variability using a fluctuating absolute threshold of hearing which improved their predictions in silence. Due to the small influence of the exact form of the absolute hearing threshold in our study, this procedure was not applied here.

E. Practical relevance

There are at least two different applications that may benefit from this modeling approach. First, this approach may be used to model sensorineural hearing loss by appropriate manipulation of the auditory preprocessing. Hence, consequences of the auditory preprocessing on speech recognition for listeners with impaired hearing can be investigated. As a long-term aim the model may serve as a tool for distinguishing between reduced speech recognition caused by impaired preprocessing or by further problems in the patient's central processing. A further long-term aim is to find processing strategies that substantially enhance the recognition performance of certain phonemes and that can be used in hearing-aids. Second, automatic speech recognizers may be improved especially for functioning in noise, if they focus on passages fitting well to their vocabulary and if they neglect outlying passages in a manner similar to that used in the weighting of the perceptual distance in this study.

V. CONCLUSIONS

(1) The microscopic approach for predicting speech intelligibility by using an auditory model as a pre-processor to a DTW speech recognizer is capable of discriminating CVC and VCV logatoms in noise.

(2) If the detector stage is assumed to be optimal by using identical templates for test signal and vocabulary, the speech discrimination performance of the model is very similar to that of human listeners. This means that the recognition of logatoms by humans can be modeled effectively by assuming a limited auditory-like preprocessing stage and a perfect speech matching process. In other words, the prediction of normal-hearing listeners' speech recognition is only possible if exactly the same stimulus is available as *a-priori* knowledge.

(3) No substantial improvement in performance of the model with *imperfect* knowledge about the speech signal was found when changing the distance measure.

(4) For the model with *perfect* knowledge about the speech signal, the Lorentzian measure is the best distance measure where outlying passages have the smallest weight compared to the other distance measures such as the Euclidean or the two-sided-exponential.

(5) Predicted recognition rates of each single phoneme are very similar to observed recognition rates but some of the observed characteristic patterns of human confusions did not occur within the predictions.

ACKNOWLEDGMENTS

We thank Birger Kollmeier for his substantial support and contribution to this work and Bernd Meyer for making available the ASR data. Thanks to Mitchell Sommers, Amy Beeston, and one anonymous reviewer who helped to greatly improve the manuscript. We would also like to thank the EU HearCom Project, the "Förderung wissenschaftlichen Nachwuchses des Landes Niedersachsen" (FwN), and SFB/TR 31 "Das aktive Gehör" (URL: http://www.uni-oldenburg.de/sfbtr31) for funding the research reported in this paper.

APPENDIX: SIGNIFICANCE OF CONFUSION MATRIX ELEMENTS

For deciding whether or not two matrix elements differ significantly, a statistical analysis has to be made. One element of a confusion matrix is given by $p=x/n$, with x denoting the number of recognitions of the phoneme specified by the column and n denoting the number of presentations specified by the row of the matrix. There are $n=50$ (VCV) and $n=80$ (CVC) presentations, respectively, of each phoneme at each SNR (i.e., each confusion matrix). Each single presentation is followed by a subjects' decision for one response alternative given in the list. Therefore, each decision

is a Bernoulli-trial with an unknown underlying probability π for the correct item and $(1-\pi)$ for all other items. Note that p is just an estimate of π. By estimating π using p, both-sided 95%-confidence intervals can be calculated based on binomial statistics (Sachs, 1999). The upper boundary is given by

$$\pi_{\text{upper}} = \frac{(x+1)F_{\text{upper}}}{n-x+(x+1)F_{\text{upper}}}, \tag{A1}$$

with $F_{\text{upper}}=F_{\{2(x+1),2(n-x)\}}$ taken from Fisher's F-distribution. The lower boundary is given by

$$\pi_{\text{lower}} = \frac{x}{x+(n-x+1)F_{\text{lower}}}, \tag{A2}$$

with $F_{\text{lower}}=F_{\{2(n-x+1),2x\}}$.

The range of confidence intervals for an observed percentage p in the speech test, i.e. $(\pi_{\text{upper}}-\pi_{\text{lower}})$, results in 4% to 22% for $n=80$ (CVC presentation) and 6% to 29% for $n=50$ (VCV presentation) whereas the wider range can be found at $p=50\%$ and the smaller range at $p=0\%$ and $p=100\%$. These confidence intervals contain the underlying probability π with a confidence of 95%. Furthermore, they offer a criterion to decide if two percentages that are statistically independent of each other differ significantly (i.e., their confidence intervals must not overlap). The precondition, statistical independence within one confusion matrix, is warranted only for two matrix elements that are not part of the same row because in this case completely different phonemes were presented to obtain the two percentages. Two elements of the same row are not independent of each other because the recognition of one phoneme affects the percentages for the other phonemes of that row. A comparison of two elements being part of the same row requires a different statistical analysis that is not discussed here. Therefore, only elements of different rows (or different confusion matrices) can be tested for difference using the methods described in this section. When comparing two different confusion matrices (e.g., observed with predicted) this problem does not occur.

[1]Breebaart *et al.* (2001) found out that a 9.4 dB SPL Gaussian noise within one gammatone filter channel just masks a sinusoid with 2 kHz frequency at absolute hearing threshold (5 dB SPL, which is about 4 dB lower). This approach was extrapolated for other audiometric frequencies.

[2]Even if very few logatoms in this corpus are forenames or may have a meaning in certain dialect regions in Germany these logatoms are not excluded in this study to preserve the very systematic composition of this speech corpus.

ANSI (**1969**). "ANSI S3.5-1969 American national standard methods for the calculation of the articulation index," Standards Secretariat, Acoustical Society of America.

ANSI (**1997**). "ANSI S3.5-1997 Methods for calculation of the speech intelligibility index," Standards Secretariat, Acoustical Society of America.

Barker, J., and Cooke, M. (**2007**). "Modelling speaker intelligibility in noise," Speech Commun. **49**, 402–417.

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.

Brand, T., and Kollmeier, B. (**2002**). "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," J. Acoust. Soc. Am. **111**, 2801–2810.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2001**). "Binaural processing model based on contralateral inhibition. I. Model structure," J. Acoust. Soc. Am. **110**, 1074–1088.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2002**). "A time-domain binaural signal detection model and its predictions for temporal resolution data," Acta. Acust. Acust. **88**, 110–112.

Chi, T. S., Gao, Y. J., Guyton, M. C., Ru, P. W., and Shamma, S. (**1999**). "Spectro-temporal modulation transfer functions and speech intelligibility," J. Acoust. Soc. Am. **106**, 2719–2732.

Christiansen, T. U., Dau, T., and Greenberg, S. (**2006**). "Spectro-temporal processing of speech—An information-theoretic framework," in *International Symposium on Hearing 2006*, Cloppenburg, edited by B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. M. Verhey (Springer, Heidelberg), pp. 517–524.

Dau, T., and Kohlrausch, A. (**1997**). "Modeling auditory processing of amplitude modulation I. Detection and masking with narrowband-carriers," J. Acoust. Soc. Am. **102**, 2893–2905.

Dau, T., Püschel, D., and Kohlrausch, A. (**1996a**). "A quantitative model of the "effective" signal processing in the auditory system: I. Model structure," J. Acoust. Soc. Am. **99**, 3615–3622.

Dau, T., Püschel, D., and Kohlrausch, A. (**1996b**). "A quantitative model of the "effective" signal processing in the auditory system: II. Simulations and measurements," J. Acoust. Soc. Am. **99**, 3623–3631.

Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. (**2001**). "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment," Audiology **40**, 148–157.

Dubno, J. R., Dirks, D. D., and Langhofer, L. R. (**1982**). "Evaluation of hearing-impaired listeners using a nonsense-syllable test. 2. Syllable recognition and consonant confusion patterns," J. Speech Hear. Res. **25**, 141–148.

Elhilali, M., Chi, T., and Shamma, S. A. (**2003**). "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Commun. **41**, 331–348.

Emiroğlu, S. (**2007**). "Timbre perception and object separation with normal and impaired hearing," Ph.D. thesis, Universität Oldenburg, Oldenburg, Germany.

Emiroğlu, S., and Kollmeier, B. (**2008**). "Timbre discrimination in normal-hearing and hearing-impaired listeners under different noise conditions," Brain Res. **1220**, 199–207.

Florentine, M., and Buus, S. (**1981**). "An excitation-pattern model for intensity discrimination," J. Acoust. Soc. Am. **70**, 1646–1654.

Ghitza, O., and Sondhi, M. M. (**1997**). "On the perceptual distance between speech segments," J. Acoust. Soc. Am. **101**, 522–529.

Hant, J. J., and Alwan, A. (**2003**). "A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise," Speech Commun. **40**, 291–313.

Hohmann, V. (**2002**). "Frequency analysis and synthesis using a gammatone filterbank," Acta. Acust. Acust. **88**, 433–442.

Holube, I., and Kollmeier, B. (**1996**). "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," J. Acoust. Soc. Am. **100**, 1703–1716.

Huber, R., and Kollmeier, B. (**2006**). "PEMO-Q—A new method for objective: Audio quality assessment using a model of auditory perception," IEEE Trans. Audio, Speech, Lang. Process. **14**, 1902–1911.

IPA (**1999**). *The Handbook of the International Phonetic Association* (Cambridge University Press, Cambridge), pp. 194–197.

Jürgens, T., Brand, T., and Kollmeier, B. (**2007**). "Modelling the human-machine gap in speech reception: Microscopic speech intelligibility prediction for normal-hearing subjects with an auditory model," in Interspeech 2007, Antwerp, Belgium, pp. 410–413.

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

Lancaster, H. O. (**1958**). "The structure of bivariate distributions," Ann. Math. Stat. **29**, 719–736.

Lippmann, R. P. (**1997**). "Speech recognition by machines and humans," Speech Commun. **22**, 1–15.

Magnusson, L. (**1996**). "Speech intelligibility index transfer functions and speech spectra for two Swedish speech recognition tests," Scand. Audiol. **25**, 59–67.

Messing, D., Delhorne, L., Bruckert, E., Braida, L., and Ghitza, O. (**2008**). "Consonant discrimination of degraded speech using an efferent-inspired closed-loop cochlear model," in Interspeech 2008, Brisbane, Australia, pp. 1052–1055.

Meyer, B., and Wesker, T. (**2006**). "A human-machine comparison in speech recognition based on a logatom corpus," in Workshop on Speech Recognition and Intrinsic Variation, Toulouse, France.
Meyer, B., Brand, T., and Kollmeier, B. (**2007a**). "Phoneme confusions in human and automatic speech recognition," in Interspeech 2007, Antwerp, Belgium, pp. 1485–1488.
Meyer, R. M., Kollmeier, B., and Brand, T. (**2007b**). "Predicting speech intelligibility in fluctuating noise," in Eighth EFAS Congress Joint Meeting with the Tenth Congress of the German Society of Audiology, Heidelberg, Germany.
Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.
Moore, B. C. J. (**2003**). "Speech processing for the hearing-impaired: Successes, failures and implications for speech mechanisms," Speech Commun. **41**, 81–91.
Pavlovic, C. V. (**1987**). "Derivation of primary parameters and procedures for use in speech intelligibility predictions," J. Acoust. Soc. Am. **82**, 413–422.
Phatak, S. A., and Allen, J. B. (**2007**). "Consonant and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am. **121**, 2312–2326.
Plomp, R. (**1976**). *Aspects of Tone Sensation* (Academic, London).
Press, W., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (**1992**). *Numerical Recipes in C* (Cambridge University Press, Cambridge).
Rhebergen, K. S., and Versfeld, N. J. (**2005**). "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," J. Acoust. Soc. Am. **117**, 2181–2192.
Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2006**). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," J. Acoust. Soc. Am. **120**, 3988–3997.
Sachs, L. (**1999**). *Angewandte Statistik* (Springer, Berlin).
Sakoe, H., and Chiba, S. (**1978**). "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-26**, 43–49.
Sroka, J. J., and Braida, L. D. (**2005**). "Human and machine consonant recognition," Speech Commun. **45**, 401–423.
Stadler, S., Leijon, A., and Hagerman, B. (**2007**). "An information theoretic approach to predict speech intelligibility for listeners with normal and impaired hearing," in Interspeech 2007, Antwerp, Belgium, pp. 389–401.
Steeneken, H. J. M., and Houtgast, T. (**1980**). "Physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, 318–326.
Turner, C. W., Souza, P. E., and Forget, L. N. (**1995**). "Use of temporal envelope cues in speech recognition by normal and hearing-impaired listeners," J. Acoust. Soc. Am. **97**, 2568–2576.
Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertins, A., and Kollmeier, B. (**2005**). "Oldenburg logatom speech corpus (OLLO) for speech recognition experiments with humans and machines," in Interspeech 2005, Lisboa, Portugal, pp. 1273–1276, freely available at http://sirius.physik.uni-oldenburg.de (Last viewed 9/11/2009).
Zurek, P. M., and Delhorne, L. A. (**1987**). "Consonant reception in noise by listeners with mild and moderate sensorineural hearing impairment," J. Acoust. Soc. Am. **82**, 1548–1559.