Predicting consonant recognition and confusions in normal-hearing listeners

Zaar, Johannes; Dau, Torsten

Published in:
Journal of the Acoustical Society of America

Link to article, DOI:
10.1121/1.4976054

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Predicting consonant recognition and confusions in normal-hearing listeners

Johannes Zaar[a] and Torsten Dau
Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark

(Received 15 April 2016; revised 20 December 2016; accepted 25 January 2017; published online 23 February 2017)

The perception of consonants in background noise has been investigated in various studies and was shown to critically depend on fine details in the stimuli. In this study, a microscopic speech perception model is proposed that represents an extension of the auditory signal processing model by Dau, Kollmeier, and Kohlrausch [(1997). J. Acoust. Soc. Am. **102**, 2892–2905]. The model was evaluated based on the extensive consonant perception data set provided by Zaar and Dau [(2015). J. Acoust. Soc. Am. **138**, 1253–1267], which was obtained with normal-hearing listeners using 15 consonant-vowel combinations mixed with white noise. Accurate predictions of the consonant recognition scores were obtained across a large range of signal-to-noise ratios. Furthermore, the model yielded convincing predictions of the consonant confusion scores, such that the predicted errors were clustered in perceptually plausible confusion groups. The large predictive power of the proposed model suggests that adaptive processes in the auditory preprocessing in combination with a cross-correlation based template-matching back end can account for some of the processes underlying consonant perception in normal-hearing listeners. The proposed model may provide a valuable framework, e.g., for investigating the effects of hearing impairment and hearing-aid signal processing on phoneme recognition.
[http://dx.doi.org/10.1121/1.4976054]

[ICB] Pages: 1051–1064

I. INTRODUCTION

The way how humans decode speech has been investigated from various perspectives. Most commonly, the percentage of correctly identified words or sentences is assessed in the presence of some acoustical interference or degradation, such as additive noise and/or reverberation. The speech reception threshold (SRT), i.e., the signal-to-noise ratio (SNR) at which, e.g., 50% correct responses are obtained, has often been used to describe the properties of the transmission channel and/or the receiver (cf. Hagerman, 1982; Nilsson *et al.*, 1994; Wagener *et al.*, 2003; Nielsen and Dau, 2009, 2011). Such speech tests provide some useful *macroscopic* information about limiting effects induced by the acoustic conditions or the global speech reception ability of listeners. However, the SRT measure is rather coarse as it reflects responses averaged across many speech tokens. Furthermore, the listeners' performance may be strongly influenced by cognitive effects as listeners can restore missing acoustic information using semantic predictability and lexical information (e.g., Miller and Licklider, 1950; Warren, 1970; Bashford *et al.*, 1992; Kashino, 2006).

Speech perception has also been studied at a more basic level using a *microscopic* approach. Several studies have reported consistent misperceptions of isolated words (e.g., Cooke, 2009; Tóth *et al.*, 2015), typically collected in conditions of speech-on-speech masking using an open response set. Such an approach excludes semantic predictability while taking the language-specific lexical possibilities for misperceptions into account. Various other studies have focused on the perception of consonants embedded in nonsense syllables (e.g., Miller and Nicely, 1955; Wang and Bilger, 1973; Phatak and Allen, 2007; Phatak *et al.*, 2008; Zaar and Dau, 2015), e.g., in the form of consonant-vowel combinations (CVs like /ba/, /ta/, etc.), typically presented in steady-state noise at various SNRs in the context of a closed response set. This approach has the advantage that (i) the contribution of higher-level semantic *and* lexical effects is eliminated due to the nonsense nature of the stimuli and that (ii) the importance of the critical[1] high-frequency speech cues is emphasized as many consonants contain high-frequency energy (cf. Li *et al.*, 2010; Li *et al.*, 2012). These aspects make consonant perception measurements an interesting tool for assessing the effects of acoustical transmission channels as well as the effects of hearing impairment and hearing-aid signal processing on fundamental speech cues.

Miller and Nicely (1955) investigated consonant perception in terms of consonant recognition and confusions, such that not only the amount of errors but also the patterns of confusions were analyzed. Their study suggested that distinct perceptual confusions among consonants may have a major effect on speech intelligibility in noise. Miller and Nicely (1955) and related studies (e.g., Wang and Bilger, 1973) used many speech tokens to represent each consonant. The obtained responses were averaged across tokens such that the data were represented as a function of consonant identity. This analysis approach was later shown to misrepresent the data since substantial perceptual differences across

[a]Electronic mail: jzaar@elektro.dtu.dk

different speech tokens of the same phonetic identity were observed (Phatak *et al.*, 2008; Singh and Allen, 2012; Toscano and Allen, 2014). Zaar and Dau (2015) employed a measure of the perceptual distance between responses obtained with CVs presented in noise to investigate the influence of various sources of perceptual variability on consonant perception. Consistent with the aforementioned studies, different speech tokens of the same phonetic identity were found to induce a large perceptual variability. Moreover, even a slight temporal shift in the steady-state masking noise waveform was shown to induce a perceptual effect when the noise waveforms were presented along with the same speech token. On the receiver side, it was found that different normal-hearing (NH) listeners with the same language background showed large perceptual differences when presented with identical stimuli, whereas the individual listeners could reproduce their responses fairly reliably in a retest. Overall, the listeners' sensitivity to fine differences in the stimuli suggests that measures of consonant perception represent a detailed descriptor of the listeners' sensory processing.

To better understand how specific effects in consonant perception are related to differences in sensory processing, computational models of speech perception may be insightful. Various *macroscopic* speech intelligibility models have been presented, which are all based on simulations of the auditory periphery in terms of frequency selectivity (e.g., ANSI, 1969, 1997; Rhebergen *et al.*, 2006), while some models also consider modulation-frequency selective processing (e.g., Houtgast *et al.*, 1980; Payton and Braida, 1999; Jørgensen and Dau, 2011; Jørgensen *et al.*, 2013). Based on the assumption that speech intelligibility is monotonically related to the speech-to-noise power ratio in the considered domain, these macroscopic models have been shown to account well for average SRTs in various acoustic conditions. Only a few modeling studies have addressed *microscopic* speech perception, where typically elaborate models of the auditory periphery have been combined with a speech recognition back end to predict nonsense syllable perception. As "blind" automatic speech recognition (ASR) systems perform much worse than human listeners in terms of phoneme recognition (e.g., Sroka and Braida, 2005; Meyer *et al.*, 2011), all microscopic speech perception models presume some kind of *a priori* information about the stimuli to reduce the gap to human recognition performance.

Messing *et al.* (2009) used a non-linear model of the auditory periphery with a feedback mechanism in combination with a simplistic template matching back end (using "frozen speech," i.e., *a priori* knowledge about the presented speech token) to predict results of a diagnostic rhyme test (DRT) obtained with NH listeners. The predictions matched the data quite well in terms of the errors as a function of phonetic attributes. Jepsen *et al.* (2014) applied a similar approach to model DRT results in hearing-impaired (HI) listeners, using a different non-linear auditory model that includes an adaptation process and a modulation filterbank (Jepsen *et al.*, 2008). However, the two studies used highly controlled *synthetic* consonant-vowel-consonant (CVC) syllables mixed with speech-shaped noise (SSN). Thus, it remained unclear to what extent these models could generalize to the less controlled case of *natural* speech stimuli.

Cooke (2006) predicted NH listeners' consonant perception obtained with natural vowel-consonant-vowel (VCV) syllables in SSN based on the spectro-temporal excitation pattern (Moore, 2003). Speech-dominated spectro-temporal "glimpses" in the speech and noise mixture were fed to a Hidden-Markov Model (HMM) based missing-data speech recognizer trained on talker-specific speech samples. While the model accounted reasonably well for the consonant-specific recognition scores, the predicted consonant confusions differed strongly from those observed in the measured data.

Holube and Kollmeier (1996) used an auditory model (Dau *et al.*, 1996) in combination with a template-matching back end to predict the recognition of CVCs in SSN in NH and HI listeners. The auditory model by Dau *et al.* (1996) consists of a linear auditory filterbank, an envelope extraction stage, a nonlinear adaptation stage, and a low-pass filter, such that the internal representation (IR) is a function of time and frequency. In order to compensate for the temporal differences in the CVC test signals and the CVC templates, Holube and Kollmeier (1996) applied a dynamic time warping (DTW) algorithm (Sakoe and Chiba, 1978) as a back end. The DTW algorithm temporally warps (i.e., locally stretches and compresses) two signals such that they ideally align in time according to some distance measure. The templates were mixed with noise at the same SNR as the test signal and the decision was based on the minimum distance between the test signal and the templates after DTW. Assuming *a priori* knowledge, the speech signal contained in the correct template was identical to the test speech token such that the distance between the two signals resulted only from the differences in the noise waveforms. The model by Holube and Kollmeier (1996), fitted to account for psychoacoustic data of the individual NH and HI listeners using the original "optimal detector" back end from Dau *et al.* (1996), was shown to predict CVC-in-noise recognition data of the individual listeners (averaged across all considered speech tokens) with good accuracy while confusions were not considered.

Focusing on consonant- and vowel-specific recognition and confusion data (measured in NH listeners using CVC and VCV syllables in SSN), Jürgens and Brand (2009) applied a modeling approach largely comparable to that of Holube and Kollmeier (1996). The difference in the model front end was mainly the use of a modulation filterbank (Dau *et al.*, 1997) instead of an envelope low pass filter, which is supported by several studies arguing that temporal modulations play a crucial role in consonant perception (e.g., Christiansen *et al.*, 2007; Gallun and Souza, 2008). In the back end, Jürgens and Brand (2009) considered different distance measures for the DTW and investigated model configurations with and without *a priori* knowledge. Their study concluded that (i) *a priori* knowledge was necessary to obtain realistic consonant recognition performance, (ii) the Lorentzian distance measure yielded the best predictions when *a priori* knowledge was used, (iii) consonant- and vowel-specific recognition scores were generally well

predicted (although the model tended to overestimate the recognition performance for many consonants at large SNRs), and (iv) the confusion predictions were inaccurate.

Thus, while the above microscopic speech perception models yielded reasonable predictions in terms of consonant-specific recognition scores, consonant confusions have not yet been predicted successfully using such stimulus-driven models. Moreover, it has been demonstrated that consonant perception depends on individual speech tokens and, to some extent, even on the specific choice of the masking noise waveforms (Zaar and Dau, 2015). The discussed models have been either evaluated with respect to the grand average recognition performance across phonemes or on phoneme-specific data that still represent averages across many speech tokens of the same type. In contrast, modeling consonant perception on a token-by-token basis has not been considered yet.

The present study considers another microscopic speech perception model that was evaluated on the basis of the extensive data set provided by Zaar and Dau (2015), obtained with 15 CVs (each represented by six speech tokens) in conditions of white masking noise at six SNRs. A similar auditory model front end as the one employed by Jürgens and Brand (2009) was used and a template-matching process was applied in the back end. In contrast to Jürgens and Brand (2009), the IR of the noise alone was subtracted from the IRs of the test signals and the templates prior to template matching (as in the models by Dau *et al.*, 1996; Dau *et al.*, 1997). Furthermore, while a DTW algorithm was applied to temporally align test signals and templates, a maximum-correlation based approach was chosen in the decision stage (cf. Dau *et al.*, 1996; Dau *et al.*, 1997), as opposed to the minimum-distance based approach by Jürgens and Brand (2009). As proposed by Dau *et al.* (1996) and Dau *et al.* (1997), a constant-variance internal noise was added in the decision stage. Finally, the speech and noise materials used in the present study (CVs in white noise) largely differed from the material used in Jürgens and Brand (2009), where CVCs and VCVs in SSN were considered. Average consonant recognition scores, consonant-specific recognition and confusion scores, as well as speech-token specific consonant recognition and confusion scores were considered to evaluate the model. Additionally, the response behavior of the listeners and the model was investigated by means of an entropy-based analysis.

II. MODEL FRAMEWORK AND EXPERIMENTAL CONDITIONS

A. Front-end processing

As in Jürgens and Brand (2009), the auditory preprocessing stages from Dau *et al.* (1997) were used. The model is shown in Fig. 1 ("auditory model"). The first stage of the model simulates the frequency selectivity of the human auditory system by means of a linear filterbank, consisting of 15 fourth-order gammatone filters with center frequencies logarithmically spaced between 315 Hz and 8 kHz. The outputs of the gammatone filters were shifted in time to time-align the peak delay of the individual gammatone filters. The second stage represents a rough approximation of the transformation of the basilar membrane vibrations into inner hair cell potentials and is realized as an envelope extraction mechanism. Each gammatone filter output signal is half-wave rectified and then filtered using a low pass filter with a cut-off frequency of 1 kHz. The third stage consists of a chain of five adaptation loops that were designed to mimic adaptive properties of the auditory periphery and to account for perceptual forward masking in human listeners (Kohlrausch and Püschel, 1988; Kohlrausch *et al.*, 1992; Dau *et al.*, 1996). For stationary signals, the adaptation loops provide an approximately logarithmic compression, whereas faster fluctuations are transformed more linearly. Therefore, the adaptation loops effectively perform an onset enhancement of the individual subband envelope representations. The time constants chosen for the five adaptation loops were $\tau_1 = 5$ ms, $\tau_2 = 50$ ms, $\tau_3 = 129$ ms, $\tau_4 = 253$ ms, and $\tau_5 = 500$ ms (taken from Dau *et al.*, 1996). The fourth stage of the model is a low-frequency modulation filterbank consisting of a third-order low pass filter with a cut-off frequency of 2 Hz in parallel with three second-order band pass

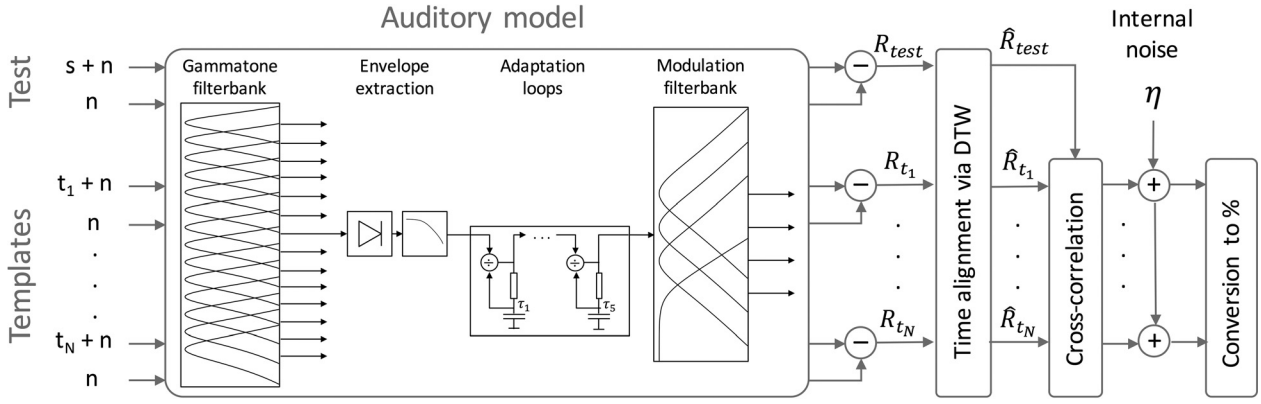

FIG. 1. Scheme of the proposed consonant perception model. For the test signal and a set of templates, the noisy speech and the noise alone were passed separately through the auditory model, consisting of a gammatone filterbank, an envelope extraction stage, a chain of adaptation loops, and a modulation filterbank. The difference between the temporal patterns of the noisy speech and the noise alone was obtained. The resulting representations of the test signal and the templates were time-aligned using a dynamic time warping (DTW) algorithm. Finally, the cross-correlation coefficients between the test signal and each template were calculated and, after addition of a constant-variance internal noise, converted to percent.

filters with a constant Q of 1 and center frequencies of 4, 8, and 16 Hz, respectively. After being fed through the adaptation loops, each subband envelope is thus further decomposed into four modulation bands. The output of the model front end obtained for any given input signal $x(t)$ is denoted as $R_x(t,f_g,f_m)$, where t denotes the temporal samples, f_g represents the gammatone filter center frequency, and f_m refers to the modulation frequency. CV speech tokens mixed with white noise were considered in this study (see Sec. II C). As in the original auditory model (Dau *et al.*, 1997), and in contrast to Holube and Kollmeier (1996) and Jürgens and Brand (2009), the noisy speech token $(s+n)$ and the noise alone (n) were separately passed through the model front end, yielding the respective temporal patterns R_{s+n} and R_n. As an input to the back end, the difference between these temporal patterns was obtained as the model's signal representation: $R_s = R_{s+n} - R_n$.

B. Speech recognition back end

The model predictions were obtained using a template-matching approach. An overview of the modeling approach is depicted in Fig. 1. In order to compare a given test signal (stimulus) with a given template, the corresponding signal representations $R_{\text{test}}(t,f_g,f_m)$ and $R_{\text{temp}}(t,f_g,f_m)$ were time aligned using a DTW algorithm as proposed by Sakoe and Chiba (1978). The DTW algorithm locally compresses and expands the time axes of two signal representations such that the temporal alignment is ideal according to the chosen distance measure. In the present study, the Euclidean distance[2] measure was used and defined as

$$D(t_i,t_j) = \sqrt{\sum_{f_g}\sum_{f_m}\left[R_{\text{test}}(t_i,f_g,f_m) - R_{\text{temp}}(t_j,f_g,f_m)\right]^2}, \tag{1}$$

where t_i and t_j denote arbitrary temporal samples. Traditionally, the chosen distance measure has also been used as a decision metric, i.e., the template showing the smallest distance to the test signal was chosen as the model response (e.g., Holube and Kollmeier, 1996; Jürgens and Brand, 2009). In the present study, however, the DTW algorithm was solely applied to obtain time aligned versions[3] of the test-signal and template representations, $\hat{R}_{\text{test}}$ and $\hat{R}_{\text{temp}}$, respectively. Inspired by the original auditory model (Dau *et al.*, 1996; Dau *et al.*, 1997), the correlation coefficient between these time-aligned representations was then calculated as the model's decision metric as

$$C\left(\hat{R}_{\text{test}},\hat{R}_{\text{temp}}\right) = \frac{\sum_{t f_g f_m}\left[\hat{R}_{\text{test}}(t,f_g,f_m) - \overline{\hat{R}_{\text{test}}}\right]\left[\hat{R}_{\text{temp}}(t,f_g,f_m) - \overline{\hat{R}_{\text{temp}}}\right]}{N_{t,g,m}\sigma_{\text{test}}\sigma_{\text{temp}}} \tag{2}$$

where $\overline{\hat{R}_{\text{test}}}$ and $\overline{\hat{R}_{\text{temp}}}$ represent the mean values and σ_{test} and σ_{temp} the standard deviations of $\hat{R}_{\text{test}}$ and $\hat{R}_{\text{temp}}$, respectively, and $N_{t,g,m}$ denotes the number of elements (number of samples × number of gammatone filters × number of modulation filters). A constant-variance Gaussian noise was added to the correlation coefficients, reflecting the listeners' uncertainty (internal noise). The variance of the noise was kept the same across experimental conditions. Eventually, the consonant corresponding to the template that yielded the largest correlation with the test signal was chosen as the model response (see Sec. II D).

C. Simulated conditions

The model was evaluated using the experimental conditions described in Zaar and Dau (2015; experiment 1). Fifteen CVs consisting of the 15 consonants /b, d, f, g, h, j, k, l, m, n, p, s, ʃ, t, v/ followed by the vowel /i/ were used whereby six recordings of each CV were taken from a Danish nonsense syllable speech material (Christiansen and Henrichsen, 2011). For each CV, three of these speech tokens were spoken by one particular male talker, the other three speech tokens were spoken by one particular female talker, amounting to a total of 90 speech tokens (15 CVs × 3 speech tokens × 2 talkers).

The speech tokens were equalized based on the peak level of an analog VU-meter simulation that responds sluggishly to the input signal (VUSOFT; Lobdell and Allen, 2007), such that they exhibited similar vowel levels while the consonant levels differed (cf. Zaar and Dau, 2015). White Gaussian noise was mixed with the speech tokens at different SNRs. SNR conditions of 12, 6, 0, −6, −12, and −15 dB were created by fixing the noise at a sound pressure level of 60 dB and adjusting the level of the speech tokens (based on the overall root-mean-square level of all speech tokens) according to the desired SNR. Each speech token was paired with one particular noise token in a given SNR condition. The noise tokens had a duration of 1 s and were faded in and out using raised cosine ramps with a duration of 50 ms. The speech tokens were mixed with the respective noise tokens such that the speech token onset was temporally positioned 400 ms after the noise onset. Eight NH native Danish listeners were presented three times with each speech token at each SNR and asked to vote for the consonant they heard. Thus, 24 responses (8 listeners × 3 repetitions) were collected per speech token and SNR, while 144 responses (8 listeners × 3 repetitions × 3 speech tokens × 2 talkers) were obtained per CV and SNR. The occurrences of responses were divided by the number of stimulus presentations to obtain the proportions of responses. The above described stimuli and the corresponding consonant perception data of Zaar and Dau (2015) were used throughout this study as inputs to the model and as reference data, respectively.

D. Simulation procedure

The same experimental stimuli the listeners had been presented with were fed to the model. While Jürgens and Brand (2009) added threshold-equalizing noise to the signals, audibility thresholds were not explicitly considered in the present study since the fixed-level masking noise was above the NH listeners' thresholds in the considered

frequency range. Each test signal (i.e., each experimental stimulus) was compared to a talker-specific template set. The speech token contained in the correct template was identical to the speech token contained in the test signal (assumption of *a priori* information); the other 14 consonants were each represented by the three available talker-specific speech tokens, such that, overall, 43 speech tokens ($1 \times 1 + 14 \times 3$) were used for the template generation. The masking noise waveforms in the test signals were the same as in the experiment. The template speech tokens were mixed with randomly generated white noise at the test-signal SNR in analogy to the stimulus generation described in Sec. II C. Five different templates were obtained from each considered speech token by mixing the speech token with five randomly generated noise waveforms. Thus, for a given test signal, the correct response alternative was represented by 5 templates (1 speech token $\times$ 5 noise tokens), whereas the other response alternatives were each represented by 15 templates (3 speech tokens $\times$ 5 noise tokens), amounting to 215 templates overall.

All test signals and templates and the corresponding noise signals were fed through the model front end, as described in Sec. II A, to obtain the respective signal representations R_{test} and R_{temp}. The signal representations were cut such that the noise-only parts at the beginning and the end were omitted and only the speech-containing portions[4] of the test signals and templates were further processed. For computational efficiency, the temporal resolution was reduced from a sampling rate of 44.1 kHz to 100 Hz by buffering R_{test} and R_{temp} into 10-ms time frames and taking the mean value across all samples within each frame. Time aligned versions of the signal representations—$\hat{R}_{\text{test}}$ and $\hat{R}_{\text{temp}}$—were obtained for each combination of test signals and templates using DTW and the correlation coefficients between them were calculated (as described in Sec. II B). As a result, correlation coefficients between each test signal and each of the respective 215 templates were obtained. Internal Gaussian noise was added to the correlation coefficients with a constant variance of $\sigma^2_{\text{int}} = 0.05$. The variance of the internal noise was chosen such that it yielded the best possible agreement of the predicted and measured grand average consonant recognition scores, i.e., the noise globally calibrated the model but did not change across SNRs, stimuli, or templates.

To convert the noisy correlation coefficients obtained for a specific test signal to proportions of responses, multiple subsets of templates were drawn from the available 215 templates. Model responses were obtained based on each template subset and finally averaged across the considered subsets. Each subset consisted of 15 templates, each representing a different response alternative (i.e., one consonant). To ensure an unbiased comparison, all feasible combinations of templates were considered as subsets. As the 14 incorrect response alternatives were each represented by 15 different templates and the correct response alternative was represented by 5 different templates (see above), the number of combinations (i.e., the number of template subsets) was $15^{14} \times 5$. For each subset, the template that showed the largest correlation with the test signal was selected as the model response. The occurrences of model responses were then divided by the number of considered template subsets to obtain the modeled proportions of responses. The procedure described above was iterated 100 times with randomly generated internal noise in each iteration and the results obtained in the individual iterations were finally averaged.

III. RESULTS AND ANALYSIS

A. Consonant recognition

Figure 2 depicts the grand average consonant recognition scores, i.e., the average recognition scores across all considered speech tokens, as a function of SNR. The open circles represent the average consonant recognition scores measured in NH listeners (Zaar and Dau, 2015). The filled black circles show the model predictions from the present study, obtained with the calibrated model (with internal noise variance $\sigma^2_{int} = 0.05$), while the small gray circles and dashed gray lines represent model predictions obtained with a range of internal noise variances ranging from $\sigma^2_{int} = 0$, i.e., no internal noise, to $\sigma^2_{int} = 0.5$. It can be observed that the predictions obtained with the calibrated model at this global level were very close to the perceptual data. This was the case for both the SRTs (data: -3 dB/predictions: -3.4 dB) and the slopes of the recognition curves. Thus, the correlation between the two curves was at ceiling (Pearson's $r = 0.998$) and the root-mean-squared error (RMSE) between them was small (RMSE = 1.68%). Regarding the role of the internal noise, the upper dashed gray lines ($\sigma^2_{int} = 0$ and $\sigma^2_{int} = 0.03$) reveal that the model overestimated consonant recognition at SNRs of 0, 6, and 12 dB when no or not enough internal noise was considered, resulting in overly steep slopes. In contrast, internal noise variances $\sigma^2_{int} > 0.05$ led to an underestimation of consonant recognition and thus to too shallow slopes. For the following figures and analyses only the calibrated model was considered.

Figure 3 shows the consonant-specific recognition scores, i.e., the consonant recognition scores averaged across speech tokens of the same phonetic identity (e.g., /bi/). The consonants are indicated in the upper left corners of the respective figure panels. Comparing the measured

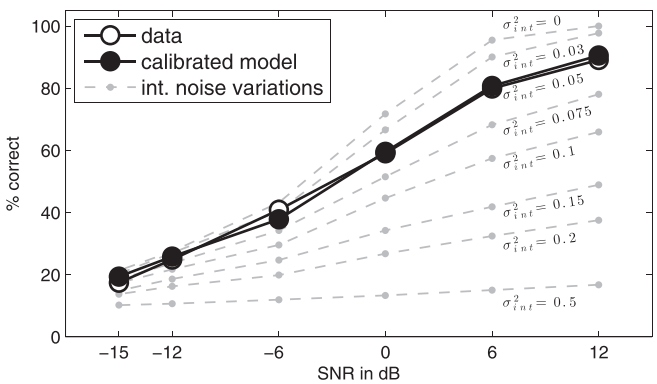

FIG. 2. Grand average consonant recognition scores in percent as a function of SNR. The open black circles represent the perceptual data and the filled black circles show the model predictions obtained with the calibrated model (internal noise variance $\sigma^2_{int} = 0.05$). The small gray circles and dashed gray lines represent model predictions obtained with a range of internal-noise variances σ^2_{int}, which are indicated next to the corresponding curves.

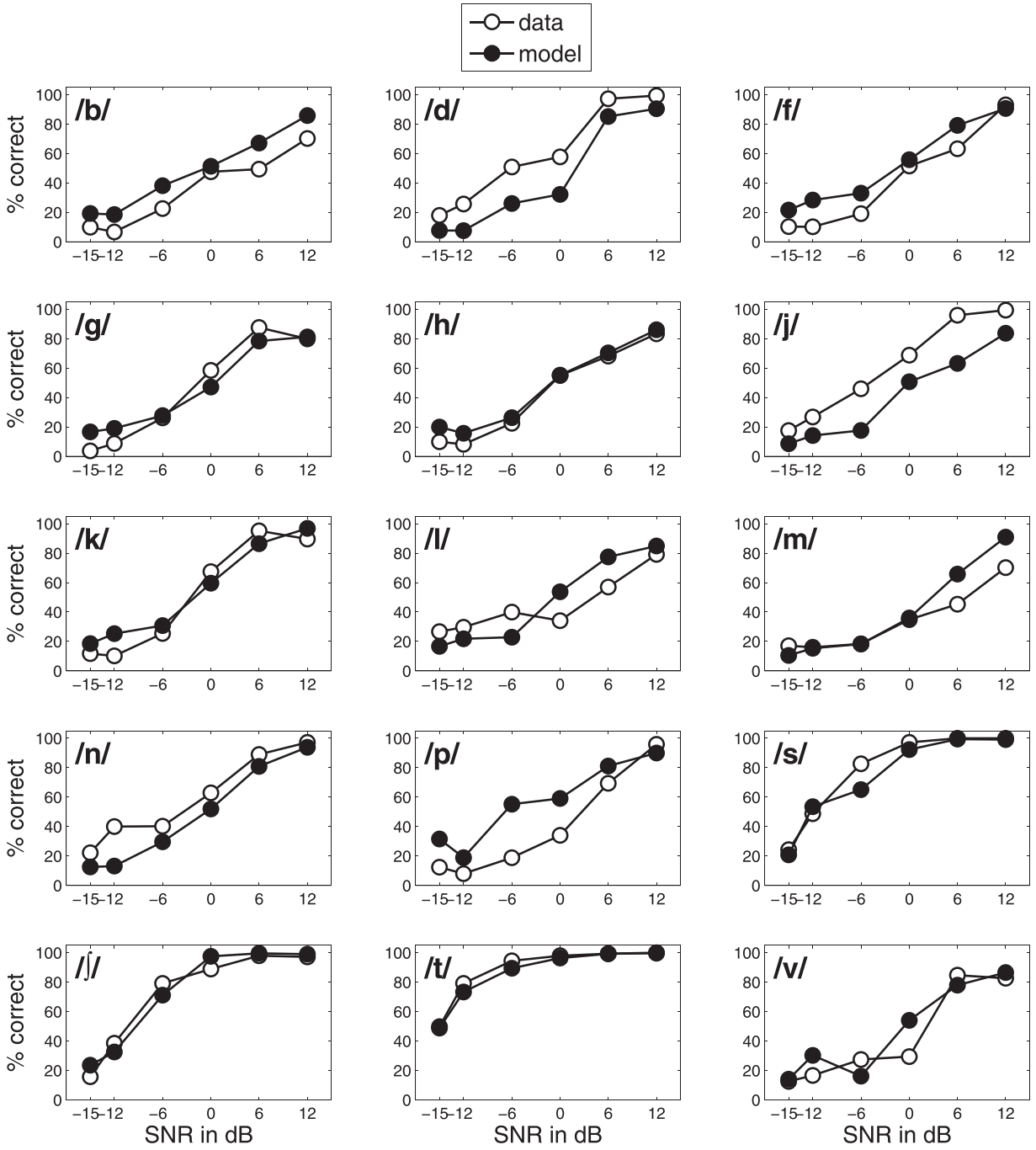

FIG. 3. Consonant-specific recognition scores in percent as a function of SNR (averaged across speech tokens of the same type). The open circles represent the perceptual data and the filled circles show the corresponding model predictions. The consonants are indicated in the upper left corners of the panels.

recognition scores (open circles) across panels, it can be observed that the individual consonants exhibited drastic differences with respect to their perceptual robustness to the influence of the masking noise. For instance, the consonant /t/ (bottom middle panel in Fig. 3) was, on average, almost perfectly recognized by listeners down to an SNR of −6 dB and still recognized about 50% of the times at −15 dB SNR. This noise robustness can also be observed for /s/ (right panel in fourth row) and /ʃ/ (bottom left panel). In contrast, some of the consonants were perceptually much more vulnerable. For example, /v/ (bottom right panel in Fig. 3) shows a recognition score of only about 80% at the large SNRs of 12 and 6 dB, followed by a sudden drop to around 30% at 0 dB SNR, from where the recognition scores approached chance-level (6.7%) performance toward lower SNRs. Equally low recognition scores can also be observed for /b/, /f/, /h/, /l/, /m/, and /p/.

The recognition scores predicted by the model are indicated as filled circles in Fig. 3. Overall, the model predictions of the consonant-specific recognition scores fit the perceptual data very well. In particular, the noise robustness of /s/, /ʃ/, and /t/ was well reflected in the predictions, as indicated by the overlap of the corresponding measured and simulated recognition curves. Furthermore, the predicted

recognition curves for most of the other consonants provided an almost exact match with the measured ones (e.g., /f, g, h, k, n, v/). In the case of /b/, /l/, /m/, and /p/, the model performed slightly better than the listeners, particularly for large SNRs. For /d/ and /j/, however, the model slightly underestimated the listeners' performance. The predicted recognition scores in these cases showed an offset across all SNRs while the predicted recognition curves were qualitatively quite similar to the measured ones.

To quantify the agreement between predictions and measurements, Pearson's r was calculated at each SNR condition between the measured and the predicted recognition scores (i) across the consonant-specific recognition scores (averaged across different speech tokens of the same type) and (ii) across the speech-token specific recognition scores. Table I summarizes the results. It can be seen that the measured and predicted recognition scores were significantly ($p < 0.05$) correlated across consonants; for SNRs of 6, 0, −6, and −12 dB the correlations were highly significant ($p < 0.01$). Correspondingly, the correlations were large particularly at medium SNRs (maximum: $r = 0.76$ at 0 dB SNR/minimum: $r = 0.55$ at 12 dB SNR). Furthermore, Table I shows that the measured and predicted recognition scores were highly significantly ($p < 0.01$) correlated even for individual speech tokens. Again, the largest correlation was observed at medium SNRs (maximum: $r = 0.57$ at −6 dB SNR/minimum: $r = 0.31$ at −15 dB SNR). As expected, the correlation coefficients across the speech-token specific recognition scores were generally lower than the correlation coefficients across the consonant-specific recognition scores. However, the p-values for the speech-token specific correlations were also lower, indicating higher significance than in the consonant-specific case. This was due to the difference in the number of data points considered for the individual correlations (15 for the consonant-specific case vs 90 for the speech-token specific case).

B. Consonant confusions

Figure 4 provides an overview of the entire measured and predicted data in terms of a confusion matrix (CM). The perceptual data and the model predictions were averaged across speech tokens of the same identity and across the six considered SNRs to obtain the CM. The vertical axis indicates the presented consonants, while the horizontal axis represents the consonants provided as response alternatives. Therefore, the full response patterns obtained for the individual consonants (consisting of the average consonant recognition as well as consonant confusion scores) are reflected in the individual rows of the CM and the average recognition scores are represented by the diagonal elements of the CM. The perceptual data and the predictions are depicted as circles, the size of which indicates the underlying proportions of responses according to the six categories shown in the figure's legend.

TABLE I. Correlation between perceptual and predicted consonant recognition scores in terms of Pearson's correlation coefficients r and the corresponding p-values. P-values indicating significant correlation ($p < 0.05$) are given in bold font. For each SNR condition, the correlation analysis was performed across consonants (left) and across individual speech tokens (right).

	Across consonants		Across speech tokens	
SNR	r	p	r	P
12 dB	0.55	**0.017**	0.35	**0.000**
6 dB	0.65	**0.004**	0.39	**0.000**
0 dB	0.76	**0.001**	0.43	**0.000**
−6 dB	0.75	**0.001**	0.57	**0.000**
−12 dB	0.75	**0.001**	0.56	**0.000**
−15 dB	0.57	**0.013**	0.31	**0.001**

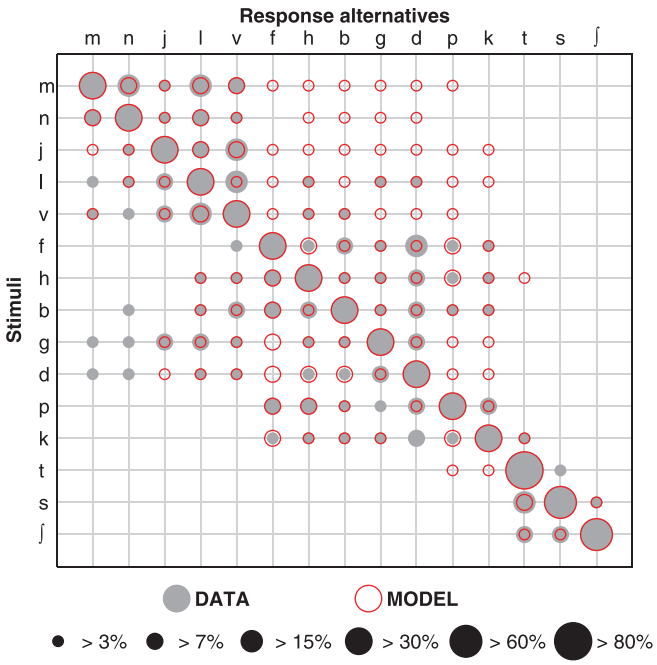

FIG. 4. Data and predictions averaged across SNR and across speech tokens of the same type, depicted as a confusion matrix. The presented consonants are shown on the vertical axis and the response alternatives on the horizontal axis. The filled gray circles represent the perceptual data while the open red circles show the model predictions. The size of the circles indicates the proportions of responses according to the six categories provided in the legend.

A complete overlap of circles indicates a large agreement between the respective measured (filled gray circles) and predicted (open red circles) average response scores. Such complete overlap can be observed along the CM's diagonal, which reflects the average consonant recognition scores. This is another view of the good agreement of measured and predicted consonant-specific recognition scores demonstrated in Table I and Fig. 3. The off-diagonal CM elements represent the average consonant confusions. Certain groups of consonants that were likely to be confused with each other (*confusion groups*) can be observed in the perceptual data (filled gray circles). Most notably, three groups can easily be identified: /m, n, j, l, v/, /f, h, b, g, d, p, k/, and /s, ʃ, t/. Additionally, there was some overlap between the first and the second group. In general, the confusion predictions of the model (open red circles) captured the measured confusions (filled gray circles) quite well, as can be seen from the overlap of the off-diagonal circles. In particular, the vast majority of the measured confusions was reflected in the predictions (70 out of 81 measured

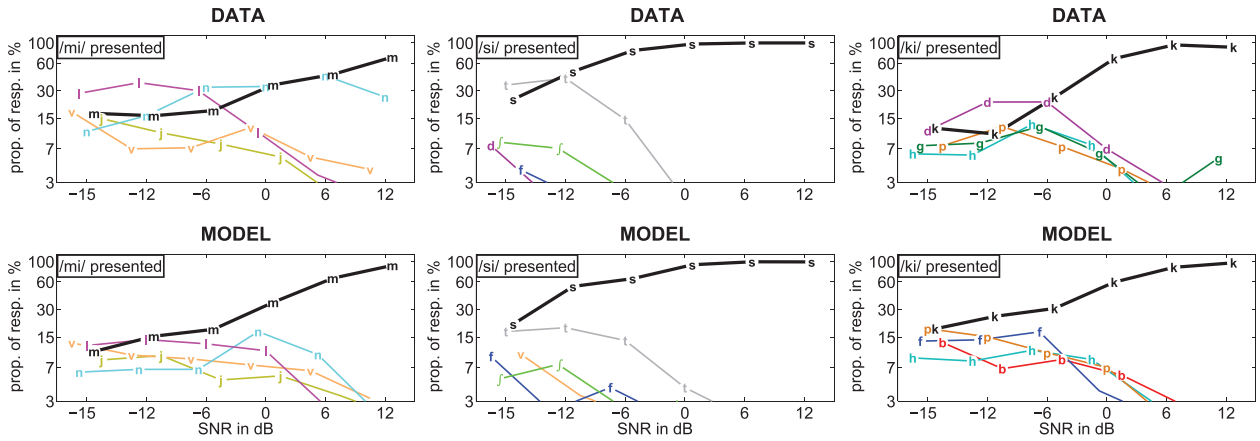

FIG. 5. Measured (top) and predicted (bottom) confusion patterns obtained for /m/ (left), /s/ (middle), and /k/ (right). The data were averaged across different speech tokens of the same type. The correct responses are indicated as thick black lines and the confusions are shown as thinner lines in different colors; the data points are labeled with the corresponding consonants. Maximally five responses are depicted for clarity, which were chosen based on their extent. A slight horizontal shift was introduced to the data for better readability. The ordinate is scaled logarithmically to emphasize the confusions.

confusions "hit" by the model according to the categories used in Fig. 4), i.e., the model's errors were, on average, very similar to the errors made by the listeners. This was also reflected in the clustering of the model predictions, which, to a large extent, followed the confusion group clustering discussed above for the perceptual data. However, the model tended to underestimate the extent of the confusions (i.e., there are many red circles that are smaller than their gray counterparts) and, instead, predicted additional confusions (e.g., /m, n, j/ confused with /f, h, b, g, d, p/) that were not reflected in the perceptual data (36 "false alarms" predicted by the model according to the categories used in Fig. 4).

Figure 5 shows three example confusion patterns (first introduced by Allen, 2005) for /m/, /s/, and /k/, respectively, each reflecting the average responses obtained with six different speech tokens. While /m/ and /s/ represent two examples with highly correlated measured and predicted confusions (Pearson's r of 0.76 and 0.96, respectively; cf. Table II), /k/ showed the smallest correlation between the measured and the predicted confusions (Pearson's r of 0.43). In the top row, the perceptual data are depicted in terms of consonant recognition (black line) and consonant confusions (colored lines) as a function of SNR. In the bottom row, the corresponding model predictions are shown. It can be observed that the model predictions captured the types of confusions made by the listeners to a large extent. /m/ was confused with /n, l, v, j/ both by the listeners and the model (left panel). /s/ (middle panel) was confused with /t, ʃ, f/ by the listeners and the model, while the fourth confusion at the lowest SNR of −15 dB differed (listeners: /d/; model: /v/). In the case of /k/, it can be seen that there still was some agreement, as the model and the listeners showed confusions with /h/ and /p/. However, the other measured confusions (/d, g/) were not reflected in the model predictions, which instead showed confusions with /f, b/. Nevertheless, the overall agreement between measured and predicted confusions was large (mean Pearson's r across consonants: 0.66; cf. Table II).

As already seen in the CM (Fig. 4), the perceptual confusions were more pronounced than the predicted ones, i.e., the listeners were more consistent in their errors than the model. This is reflected in the generally lower confusion scores obtained in the model predictions as compared to the perceptual data. For instance, in the case of /m/ (top left panel), the listeners showed a very pronounced confusion with /n/, which reached up to 44% at 6 dB SNR. In the model predictions (bottom left panel), however, the maximum confusion with /n/ reached only 17% (at 0 dB SNR). Similar underestimations of the confusions can be observed for the consonant /s/ (middle panel), as well as for many other

TABLE II. Correlation between perceptual and predicted consonant confusion scores as a function of the presented consonant (only obtained if the overall error $P_e > 20\%$). The Pearson's correlation coefficients r and the corresponding p-values were obtained across the response alternatives (excluding the recognition scores) based on the consonant-specific and on the speech-token specific across-SNR average data, respectively. The speech-token specific correlation results were then averaged across the different speech tokens of the same type (averages $\bar{r}$ and $\bar{p}$). P-values indicating significant confusion correlation ($p < 0.05$) are given in bold font. The rightmost column additionally contains the number N_s of tokens showing significant confusion correlation ($p < 0.05$) and the number N_c of considered tokens (with error $P_e > 20\%$). The consonants are ordered as in Fig. 4.

	Consonant-specific data		Speech-token specific data	
Consonant	r	p	$\bar{r}$	$\bar{p}$ (N_s/N_c)
/m/	0.76	**0.001**	0.56	**0.045** (4/6)
/n/	0.76	**0.001**	0.58	**0.042** (5/6)
/j/	0.68	**0.004**	0.51	0.095 (4/6)
/l/	0.45	0.053	0.24	0.332 (2/6)
/v/	0.60	**0.012**	0.37	0.222 (2/6)
/f/	0.60	**0.011**	0.42	0.098 (3/6)
/h/	0.69	**0.003**	0.45	0.117 (2/6)
/b/	0.65	**0.006**	0.52	0.055 (3/6)
/g/	0.60	**0.012**	0.55	**0.048** (4/6)
/d/	0.49	**0.038**	0.31	0.217 (2/6)
/p/	0.82	**0.000**	0.55	**0.047** (4/6)
/k/	0.43	0.060	0.33	0.176 (2/6)
/t/	N/A	N/A	−0.02	0.520 (0/2)
/s/	0.96	**0.000**	0.89	**0.000** (4/4)
/ʃ/	0.80	**0.000**	0.54	0.058 (3/5)

consonants that exhibited large perceptual confusions (not shown here).

To evaluate the significance of the observed agreement between the confusions in the perceptual data and in the model predictions, Pearson's r was calculated between the measured and predicted across-SNR average response patterns using (i) consonant-specific data (i.e., data averaged across different speech tokens of the same type) and (ii) speech-token specific data. Only the erroneous responses obtained for each CV/speech token (i.e., only the off-diagonal elements of the CM) were correlated; the recognition scores (on-diagonal elements of the CM), which would otherwise strongly dominate the correlations, were excluded in order to evaluate the qualitative agreement of the measured and predicted confusions irrespective of the recognition score agreement. This *confusion correlation* was only taken into account if the cumulative error P_e (i.e., the sum of all perceptual confusions averaged across SNR) exceeded 20%.

The left part of Table II summarizes the results obtained with the consonant-specific data in terms of a correlation coefficient r and a corresponding p-value for each stimulus consonant. The analysis revealed that the predicted confusions were strongly correlated with the measured confusions when considered at the consonant level (maximum: $r = 0.96$ for /s/; minimum: $r = 0.43$ for /k/; average: $r_{avg} = 0.66$). Almost all (12 out of 15) consonant-specific confusion correlations were significant ($p < 0.05$, in bold font), except for /l/ and /k/, which exhibited p-values just above 0.05. For /t/, no correlation was obtained as the error was too small ($P_e \leq 20\%$).

For the speech-token specific case, correlation coefficients and p-values were obtained for each of the 90 speech tokens. For the sake of compactness, the right side of Table II shows a collapsed version of the results obtained with the speech-token specific data in terms of the average correlation coefficients $\bar{r}$ and the average p-values $\bar{p}$ for each stimulus consonant (i.e., averaged across speech tokens of the same type). Additionally, the number of significantly correlated confusion patterns ($p < 0.05$), N_s, and the number of considered speech tokens (with $P_e > 20\%$), N_c, are provided in the rightmost column of Table II. The speech-token specific confusion correlation analysis revealed that the confusion correlations were significant only for 43 of the 83 eligible speech tokens (7 of the 90 speech tokens showed $P_e \leq 20\%$ and were thus not considered). The maximum average confusion correlation at the speech-token level was $\bar{r} = 0.89$ for /ʃ/. All other correlations were much smaller, with a minimum at $\bar{r} = -0.02$ for /t/. The average confusion correlation coefficient across all considered 83 speech tokens was $\bar{r}_{avg} = 0.47$.

In addition to the confusion correlation analysis of the across-SNR average data described above, the consonant-specific and speech-token specific confusion correlations were also evaluated for the individual SNR conditions. The left side of Table III shows the average correlation coefficients and p-values obtained based on the consonant-specific data. The number N_s of consonants exhibiting significant confusion correlation ($p < 0.05$) and the number N_c of considered consonants (with $P_e > 20\%$) are given in parentheses. It can be observed that the model captured most of the measured confusions well at the consonant- and SNR-specific level. Average confusion correlations ranged between 0.44 and 0.66 and the highest correlation values were obtained for SNRs of 0 and 6 dB. The model showed significant confusion correlations for almost all (19 out of 22) considered consonants at SNRs $\geq$ 0 dB and for more than half (24 out of 43) of the considered consonants at negative SNRs. When considering the speech-token specific data per SNR (right side of Table III), the average confusion correlations were substantially lower, ranging from 0.24 to 0.44. The largest average correlations were again found for SNRs $\geq$ 0 dB, with significant confusion correlations obtained for about half (57 out of 113) of the considered speech tokens. For negative SNRs, the confusions were significantly correlated for only 30% (75 out of 250) of the considered speech tokens. This substantial decrease of the model performance at the level of individual speech tokens and SNRs was probably caused by the extremely low number of observations[5] considered in this case, which resulted in noisy reference data.

TABLE III. Correlation between perceptual and predicted consonant confusion scores as a function of SNR (only obtained if the overall error $P_e > 20\%$). For each SNR condition, the Pearson's correlation coefficients r and the corresponding p-values were obtained across the response alternatives (excluding the recognition scores) based on the consonant-specific and on the speech-token specific data, respectively. The SNR-specific results were then averaged across the different consonants and the different speech tokens, respectively (averages $\bar{r}$ and $\bar{p}$). P-values indicating significant confusion correlation ($p < 0.05$) are given in bold font. The p-values are accompanied by the number N_s of consonants/tokens showing significant confusion correlation ($p < 0.05$) and the number N_c of considered consonants/tokens with error $P_e > 20\%$ (maximally 15 consonants/90 speech tokens).

	Consonant-specific data		Speech-token specific data	
SNR	$\bar{r}$	$\bar{p}$ (N_s/N_c)	$\bar{r}$	$\bar{p}$ (N_s/N_c)
12 dB	0.47	0.097 (2/4)	0.44	0.156 (11/20)
6 dB	0.66	**0.014** (6/6)	0.37	0.226 (14/33)
0 dB	0.66	**0.019** (11/12)	0.43	0.165 (32/60)
−6 dB	0.44	0.118 (7/13)	0.27	0.258 (21/77)
−12 dB	0.45	0.157 (9/15)	0.26	0.283 (27/85)
−15 dB	0.49	0.104 (8/15)	0.24	0.294 (27/88)

C. Entropy-based analysis

The above analysis demonstrated that while the model mostly accounted for the types of measured confusions, it showed a tendency to underestimate the amount of these confusions and, instead, additionally selected other confusions that were not reflected in the perceptual data. This suggests that the model responded more randomly than the listeners. To analyze the overall response behavior of the listeners and the model in terms of the randomness of the responses, the entropy of responses was calculated (cf. Miller and Nicely, 1955; Phatak *et al.*, 2008; Zaar and Dau, 2015). In particular, the *normalized entropy* for a given response vector $\mathbf{p} = [p_1, p_2, \ldots, p_R]$, with $p_1, \ldots, p_R$ denoting

the proportions of responses for the individual R response alternatives, was defined as

$$\mathcal{H}_{\text{norm}}(\mathbf{p}) = \frac{100\%}{\log_2(R)} \sum_{i=1}^{R} \mathrm{p}_i \log_2\left(\frac{1}{\mathrm{p}_i}\right), \quad \forall\, \mathrm{p}_i > 0, \qquad (3)$$

with $\log_2(R)$ representing the theoretical entropy maximum. The normalized entropy is therefore confined to the interval [0%, 100%]. When the randomness in the response vector is minimal, i.e., one element has a value of 1 and the other elements are 0, the normalized entropy is 0%. When the randomness in the response vector is maximal, i.e., all elements have the same value of $1/R$, the normalized entropy is 100%. The normalized entropy was calculated per SNR condition (i) for each response vector in the consonant-specific perceptual data and predictions and (ii) for each response vector in the speech-token specific perceptual data and predictions and, finally, averaged across consonants and speech tokens, respectively.

Figure 6 shows the normalized entropy obtained from the perceptual data (white bars) and from the model predictions (gray bars) as a function of SNR for the consonant-specific case (left panel) and for the speech-token specific case (right panel). The entropy generally increased with decreasing SNR as the task became more challenging and the consonant percept became more uncertain due to the increased masking effect of the noise, such that more errors and less systematic errors occurred. Furthermore, the entropy in the consonant-specific perceptual data (left panel, white bars) was around 10% larger than the entropy in the speech-token specific perceptual data (right panel, white bars), except at the largest SNR of 12 dB (5% difference). This indicates that averaging across speech tokens of the same type increases the randomness in the responses, implying perceptual differences across the considered speech tokens. This effect has already been shown for the considered data set on a listener-by-listener basis (Zaar and Dau, 2015) and is here confirmed for the across-listener average data, highlighting the importance of considering the data (and predictions) at the speech-token level.

Regarding the comparison between the perceptual data and the model predictions, the entropy analysis revealed that the model predictions showed a larger entropy than the perceptual data. This was the case both for the entropy analysis at the consonant level (left panel of Fig. 6, gray bars vs white bars) and at the speech-token level (right panel, gray bars vs white bars), with differences of up to 13% in both cases. Thus, the entropy-based analysis showed that the model's response behavior was indeed more random than that of the listener panel.

IV. DISCUSSION

A. Relation to other studies

The model proposed in the present study represents an extension of the auditory detection model by Dau *et al.* (1997) toward predicting microscopic speech perception data. The main references for comparison of the model performance are the related modeling work of Jürgens and Brand (2009), which partly inspired the present study, and the Glimpse-model approach by Cooke (2006). However, it should be noted that these models were evaluated on different stimuli and data, such that a direct comparison is difficult. In particular, Jürgens and Brand (2009) used VCVs in steady-state SSN and Cooke (2006) employed VCVs in N-talker babble modulated SSN, while the present study used CVs in steady-state white noise (cf. Zaar and Dau, 2015). In terms of the grand average consonant recognition as a function of SNR, the proposed model showed an almost perfect fit with the perceptual data, whereas the model by Jürgens and Brand (2009) showed an overly steep recognition curve in their study (see their Fig. 3); this is mainly attributable to the calibration of the proposed model using internal noise (as shown in Fig. 2, see also Sec. IV B), which had not been performed by Jürgens and Brand (2009). Cooke (2006) only considered one SNR condition, such that no comparison is feasible here. Regarding consonant-specific recognition scores, Jürgens and Brand (2009) showed a good agreement between their perceptual data and the corresponding predictions at medium to low SNRs, whereas their model predicted perfect recognition irrespective of the considered consonant at large SNRs, which was not reflected in their perceptual data (see their Fig. 4). Cooke (2006) obtained reasonable predictions of the consonant-specific trends in the recognition

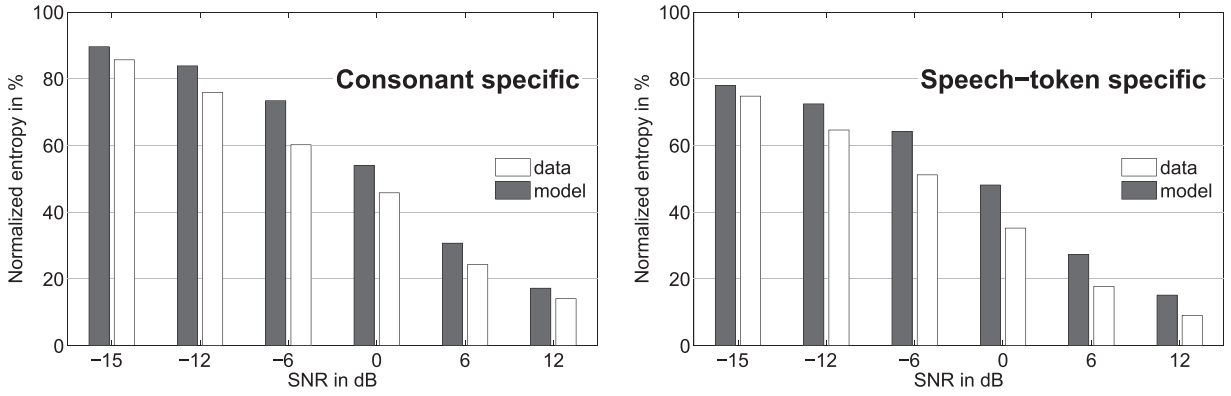

FIG. 6. Normalized entropy in percent as a function of SNR calculated from the perceptual data (white bars) and the model predictions (gray bars). Left: normalized entropy obtained from consonant-specific data and predictions; right: normalized entropy obtained from speech-token specific data and predictions. The normalized entropy was calculated for each consonant/speech token and SNR and then averaged across consonants/speech tokens.

scores for the considered SNR of −6 dB (see Cooke's Fig. 10). The model presented in the current study, however, provided significantly correlated recognition scores across consonants at *all* considered SNR conditions, including large positive SNRs of 6 and 12 dB. Furthermore, the proposed model yielded highly significantly correlated recognition score predictions even at the speech-token level, which has so far not been reported in the related literature. Finally, while Jürgens and Brand (2009) and Cooke (2006) concluded that their respective models did not account well for consonant confusions, the present study demonstrated that the proposed model predicted the perceptual consonant confusions to a large extent (at the consonant-specific level).

B. Significance of the model components

During the development of the proposed model, many decisions were taken regarding the model design. This section lays out the reasons for including the individual model components and how they influence the predictions.

The auditory model used as a front end (Dau *et al.*, 1997) was adapted for consonant perception modeling in a similar way as in Jürgens and Brand (2009). The low-frequency bands (between 50 and 300 Hz), typically considered in the gammatone filterbank, were omitted in order to mitigate the effect of differences in the low-frequency vowel portions of the stimuli and the templates, which may otherwise result in undesired effects that are independent of the consonant cues (e.g., prediction biases based on vowel-portion similarity[6]). The envelope extraction stage and the adaptation loops were parametrized as suggested by Dau *et al.* (1997). The onset enhancement performed by the adaptation loops provided realistic predictions as, e.g., the onset of the high-frequency frication noise of an /s/ was enhanced such that it became more similar to the high-frequency burst of a /t/, which led to a perceptually plausible confusion at low SNRs (see Fig. 5, middle panels). Finally, four low-frequency modulation filters were applied, as also proposed by Jürgens and Brand (2009). It should be noted that simulations obtained using a simple low-pass filter with a cut-off frequency of 8 Hz (Dau *et al.*, 1996) instead of a modulation filterbank led to comparably accurate results. However, the modulation-filterbank model is expected to generalize to a broader range of conditions as (i) the corresponding Dau *et al.* (1997) model accounts for more psychoacoustic conditions than the Dau *et al.* (1996) model and (ii) modulation-domain based macroscopic speech intelligibility models (e.g., Houtgast *et al.*, 1980; Jørgensen *et al.*, 2013) have been shown to account for a large variety of acoustic conditions.

Similar to the model of Jürgens and Brand (2009), the proposed model assumes *a priori* knowledge about the speech token contained in the test signal. Thus, the only difference between the test signal and the "correct" template was induced by the different masking noise waveforms. Without such *a priori* knowledge regarding the test speech token, the model's recognition scores were substantially lower than the measured ones. This was expected given the well-known gap between human and machine speech recognition performance (e.g., Meyer *et al.*, 2011) and the simplistic nature of the applied speech recognition back end. While Jürgens and Brand (2009) directly fed the outputs of the model front end obtained with the noisy speech tokens to the back end, the present study followed the original model from Dau *et al.* (1997) in that the difference between the front-end outputs obtained with the noisy speech and the noise alone was considered in the back end. This assumption of *a priori* knowledge about the masking noise was necessary to correctly predict the robustness of high-frequency cues (observed in the perceptual data for /s, ʃ, t/). In contrast, Jürgens and Brand (2009) could partly predict the robustness of high-frequency cues (/t, s, ts, ʃ/, see their Fig. 4) without this assumption. However, they used masking noise with a speech-shaped spectrum (sloping down toward high frequencies), such that the masking in the relevant high-frequency region was much less effective than in the present study, where white masking noise with a flat spectrum was employed. Thus, it can be concluded that if all the relevant consonant cues are masked to a comparable extent, the assumption of *a priori* knowledge about the masking noise appears to be necessary for realistic predictions, at least when using the auditory model of Dau *et al.* (1997) as a front end. The need for such a mechanism in the model is consistent with the results from a study by Mesgarani *et al.* (2014), which showed that spectrograms reconstructed from neural representations of noisy phonemes measured in ferret primary auditory cortex were more similar to the clean phonemes than to the noisy ones. This implies the existence of a de-noising mechanism at higher stages of auditory processing, which the auditory model considered in the present study does not capture. Using *a priori* knowledge about the noise may thus be considered as a simplistic way of simulating a de-noising mechanism.

The model's decision was based on the maximum cross-correlation (as in Dau *et al.*, 1997; see also Gallun and Souza, 2008) of the time-aligned IRs of the test signal and the templates, as opposed to the minimum distance used by Jürgens and Brand (2009). The cross-correlation has the advantage that it is insensitive to level differences (i.e., solely describes *covariation*), which may be more closely related to the perceptual decision-making process than any distance measures (be it Euclidean or Lorentzian distance), which are typically sensitive to level differences. An earlier distance-based version of the model indeed yielded less convincing predictions of the perceptual data, partly due to biases that were presumably induced by this level sensitivity. A similarly biased behavior can be observed in the Jürgens and Brand (2009) predictions (see their Fig. 6, panel 2). The correlation-based back end alleviated this problem to a large extent and, thus, yielded realistic predictions in terms of consonant recognition and confusion scores.

Finally, the constant-variance internal noise in the model's decision stage (representing the listeners' uncertainty, cf. Dau *et al.*, 1997) provided a realistic amount of uncertainty at medium to large SNRs, where the predicted recognition scores otherwise exceeded the measured ones, leading to overly steep recognition curves (see upper gray curve in Fig. 2). This result has also been reported by Jürgens and Brand (2009), who did not include an explicit calibration

mechanism in their model. Although the internal noise affected the model predictions differently at different SNRs (cf. Fig. 2), the internal noise used in the present study merely calibrated the model as a whole, i.e., it did not change across SNRs, stimuli, or templates. The entropy-based analysis showed that the model responded slightly more randomly than the listeners did. It might seem intuitive to reduce the internal-noise variance in order to mitigate this mismatch; however, this is not feasible as it would considerably worsen the model's prediction accuracy with respect to the consonant recognition scores.

C. Relation to data-driven approaches

The current study presents a *stimulus-driven* modeling approach, which is based on the acoustical stimuli and uses only a minimum amount of knowledge about the experimental data (for calibration). However, consonant perception data may also be predicted using a fundamentally different, *data-driven*, approach that is solely based on data obtained from another experiment. In order to compare the proposed model to such a data-driven approach, it was investigated to what extent a subset of the data set could predict the remaining data. In particular, the data set was split into responses obtained with a randomly chosen set of speech tokens (one for each considered CV). These reference data were compared to the remaining test data in terms of consonant recognition and confusions using the same measures as described in Sec. III. The same comparison was conducted between the test data and the corresponding stimulus-driven model predictions. The procedure was iterated 100 times with random splits of the data set and the derived measures were then averaged across iterations.

The analysis revealed that the grand average recognition score predictions obtained using the data-driven approach (Pearson's $r = 0.98$, RMSE $= 8.5\%$) described the data substantially less accurately than the ones obtained using the proposed stimulus-driven model (Pearson's $r = 0.99$, RMSE $= 4.1\%$). A similar trend was, on average, observed regarding the correlation between the predicted and measured consonant-specific recognition scores across consonants (data-driven: $r = 0.51$; stimulus-driven: $r = 0.61$), as well as in terms of the correlation between the predicted and measured speech-token specific recognition scores across speech tokens (data-driven: $r = 0.42$; stimulus-driven: $r = 0.44$). Regarding consonant confusions, the data-driven approach yielded, on average, higher correlations between the predicted and measured consonant-specific confusions (data-driven: $r = 0.73$; stimulus-driven: $r = 0.56$) and between the predicted and measured speech-token specific confusions (data-driven: $r = 0.67$; stimulus-driven: $r = 0.46$) as compared to the proposed model. Overall, the proposed stimulus-driven model outperformed the data-driven approach in terms of recognition score predictions, whereas the consonant confusions were better captured by the data-driven approach. However, since the reference data for the data-driven predictions were obtained from the same experiment as the test data, the corresponding stimuli were very similar and the listeners even the same. A realistic data-driven prediction, which would be based on data from one experiment and applied to data from another experiment, might yield a lower predictive power due to larger differences between reference data and test data.

D. Limitations of the approach

Despite its large predictive power for the considered data/stimuli, the consonant perception model proposed in the present study needs to be tested for generalizability using other data sets that differ with respect to the speech tokens (e.g., VCVs instead of CVs), the native language of the talkers and listeners (e.g., English instead of Danish), and/or the noise type (e.g., SSN instead of white noise). Furthermore, as all stimuli were above NH audibility thresholds in the considered frequency bands, no audibility thresholds were considered in the model. Therefore, the model is bound to fail for partly or fully inaudible stimuli due to low presentation levels or hearing impairment. This could be overcome by adding threshold-simulating noise (cf. Jürgens and Brand, 2009; Jürgens *et al.*, 2014) or by excluding the frequency bands below threshold from further processing (cf. Jørgensen and Dau, 2011). Moreover, it has been shown in Zaar and Dau (2015), based on the data set considered in the present study, that different NH listeners with the same language background can exhibit large perceptual differences for identical stimuli. The current study, however, focused on the across-listener average data, thus neglecting the across-listener perceptual variability. The proposed model has, in its current form, no means of explaining such listener-specific effects, which may be attributable to individual biases or supra-threshold processing deficits that were not captured by the audiometric test.

E. Perspectives

The most common acoustic condition that has been considered in consonant perception studies is additive stationary noise (e.g., Miller and Nicely, 1955; Wang and Bilger, 1973; Phatak and Allen, 2007; Phatak *et al.*, 2008; Zaar and Dau, 2015). While this condition has provided valuable insights in the cues underlying consonant perception, it does not reflect realistic acoustic scenarios, in which most competing sound sources are strongly modulated and reverberation is typically present. An experimental investigation of consonant perception in such conditions and a subsequent evaluation of the proposed model's predictive power for the corresponding data may therefore be a crucial next step.

The present study focused on modeling consonant perception data obtained with NH listeners. However, consonant perception measurements may be particularly insightful when used as a tool to identify specific problems experienced by HI listeners. To better understand the cause of these problems, a version of the model that is conceptually capable of explaining effects of hearing impairment may be useful. To that end, sensitivity, compression, and frequency selectivity should be adjustable in the model front end. Furthermore, a model version that simulates the effects of hearing-aid signal processing in combination with the effects of certain types of hearing impairment may be a powerful

tool for parametrizing hearing aid algorithms. A comparable model extension may be conceived for simulating the effects of cochlear-implant phoneme transduction and adjusting the corresponding algorithms (e.g., regarding channel selection).

The proposed model predicts consonant perception from an auditory modeling perspective, i.e., using *a priori* information where necessary to predict the data. A "blind" model that bases its predictions only on the stimulus, just like listeners give their responses solely based on the stimulus, would represent a more elegant approach. Such a model requires a massive ASR back end that reaches human performance, which has so far not been feasible (Meyer *et al.*, 2011). However, recent advances in ASR using HMMs in combination with Deep Neural Networks (DNNs, e.g., Hinton *et al.*, 2012; Dahl *et al.*, 2012) suggest that the gap between human and machine speech recognition is decreasing substantially. When blind ASR-based models become technically feasible, the present study may serve as a reference with respect to the front end features that should be considered to obtain realistic predictions. Furthermore, the reported predictive power of the assumption of *a priori* knowledge about the masking noise motivates the use of suitable source separation algorithms prior to the speech recognition process.

V. SUMMARY AND CONCLUSIONS

A consonant perception model was presented and evaluated with respect to consonant recognition and consonant confusions at different levels of detail. The model consists of an auditory modeling front end in combination with a correlation-based template-matching back end and represents an extension of the auditory processing model by Dau *et al.* (1997) toward predicting microscopic speech perception data. The model was evaluated based on the extensive CV-in-noise data from Zaar and Dau (2015), obtained with NH listeners. Overall, a good agreement between the perceptual data and the model predictions was demonstrated. The measured grand average consonant recognition scores as a function of SNR were almost perfectly accounted for by the model. Furthermore, the predicted consonant-specific recognition scores were highly correlated with the measured ones. Even at the speech-token level, large correlations between the predicted and the perceptual recognition scores were obtained. Regarding consonant confusions, the model predictions showed a strong similarity with the measured confusions at the consonant-specific level. However, the model tended to underestimate the extent of the main confusions in this scenario and showed only partially satisfactory confusion predictions at the speech-token level. It was shown in an additional entropy-based analysis that the model generally responded slightly more randomly than the listener panel did, which explains the observed shortcomings.

Overall, the large predictive power of the proposed model suggests that adaptive processes in the auditory preprocessing in combination with a cross-correlation based template-matching back end functionally account for some of the processes underlying consonant perception in normal-hearing listeners. The modeling framework may serve as a normal-hearing baseline for future microscopic models of speech perception that can account for effects of hearing-impairment and hearing-aid signal processing on phoneme perception.

ACKNOWLEDGMENTS

We would like to thank two anonymous reviewers for their helpful and supportive comments. This research was funded with support from the European Commission under Contract No. FP7-PEOPLE-2011-290000.

[1]Many hearing-impaired listeners suffer from a loss of sensitivity at high-frequencies, which affects their speech perception but is difficult to measure using macroscopic tests with meaningful, low-frequency dominated speech.

[2]The Euclidean distance was used for DTW as it yielded more plausible time alignment results as compared to the Lorentzian distance suggested by Jürgens and Brand (2009).

[3]The DTW algorithm from Sakoe and Chiba (1978) was applied without any path limitations, such that any local time-axis warping was in principle allowed.

[4]The start and end times of the speech-containing portions were defined as the first and last sample of the corresponding clean speech token's power (in dB) that were less than 40 dB below the speech token's power maximum.

[5]Twenty-four observations were available per speech token and SNR condition; in case of the error P_e just exceeding the 20% threshold, the considered confusion patterns therefore consisted of only five observations.

[6]Since the low-frequency energy of the CVs is large compared to the mid- and high-frequency bands but does not contribute substantially to consonant perception, slight differences in the vowel pronunciation can induce biases that are independent of consonant-cue similarity and thus detrimental to the predictive power of the model.

Allen, J. B. (**2005**). "Consonant recognition and the articulation index," J. Acoust. Soc. Am. **117**(4), 2212–2223.

ANSI (**1969**). S3.5, *American National Standard Methods for the Calculation of the Articulation Index* (Acoustical Society of America, New York).

ANSI (**1997**). S3.5, *Methods for the Calculation of the Speech Intelligibility Index* (Acoustical Society of America, New York).

Bashford, J. A., Jr., Riener, K. R., and Warren, R. M. (**1992**). "Increasing the intelligibility of speech through multiple phonemic restorations," Percept. Psychophys. **51**(3), 211–217.

Christiansen, T. U., Dau, T., and Greenberg, S. (**2007**). "Spectro-temporal processing of speech – An information-theoretic framework," in *Hearing – From Sensory Processing to Perception* (Springer, Berlin, Germany), pp. 517–523.

Christiansen, T. U., and Juel Henrichsen, P. (**2011**). "Objective evaluation of consonant-vowel pairs produced by native speakers of Danish," in *Forum Acusticum*, 2011.

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

Cooke, M. (**2009**). "Discovering consistent word confusions in noise," in *Proceedings of Interspeech*, pp. 1887–1890.

Dahl, G. E., Yu, D., Deng, L., and Acero, A. (**2012**). "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," IEEE Trans. Audio Speech Language Process. **20**(1), 30–42.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997**). "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," J. Acoust. Soc. Am. **102**(5), 2892–2905.

Dau, T., Püschel, D., and Kohlrausch, A. (**1996**). "A quantitative model of the 'effective' signal processing in the auditory system. I. Model structure," J. Acoust. Soc. Am. **99**(6), 3615–3622.

Gallun, F., and Souza, P. (**2008**). "Exploring the role of the modulation spectrum in phoneme recognition," Ear Hear. **29**(5), 800–813.

Hagerman, B. (**1982**). "Sentences for testing speech intelligibility in noise," Scand. Audiol. **11**, 79–87.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B.

(**2012**). "Deep neural networks for acoustic modeling in speech recognition – The shared views of four research groups," IEEE Signal Process. Mag. **29**, 82–97.
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**(3), 1703–1716.
Houtgast, T., Steeneken, H., and Plomp, R. (**1980**). "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics," Acustica **46**(1), 60–72.
Jepsen, M. L., Dau, T., and Ghitza, O. (**2014**). "Refining a model of hearing impairment using speech psychophysics," J. Acoust. Soc. Am. **135**(4), EL179–EL185.
Jepsen, M. L., Ewert, S. D., and Dau, T. (**2008**). "A computational model of human auditory signal processing and perception," J. Acoust. Soc. Am. **124**(1), 422–438.
Jørgensen, S., and Dau, T. (**2011**). "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," J. Acoust. Soc. Am. **130**(3), 1475–1487.
Jørgensen, S., Ewert, S. D., and Dau, T. (**2013**). "A multi-resolution envelope-power based model for speech intelligibility," J. Acoust. Soc. Am. **134**(1), 436–446.
Jürgens, T., and Brand, T. (**2009**). "Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model," J. Acoust. Soc. Am. **126**(5), 2635–2648.
Jürgens, T., Ewert, S. D., Kollmeier, B., and Brand, T. (**2014**). "Prediction of consonant recognition in quiet for listeners with normal and impaired hearing using an auditory model," J. Acoust. Soc. Am. **135**(3), 1506–1517.
Kashino, M. (**2006**). "Phonemic restoration: The brain creates missing speech sounds," Acoust. Sci. Tech. **27**(6), 318–321.
Kohlrausch, A., and Püschel, D. (**1988**). "Interrelations between a psychoacoustical model of temporal effects in hearing and neurophysiological observations," in *Sense Organs*, edited by N. Elsner and F. G. Barth (Thieme, Stuttgart, Germany), p. 39.
Kohlrausch, A., Püschel, D., and Alphei, H. (**1992**). "Temporal resolution and modulation analysis in models of the auditory system," in *The Auditory Processing of Speech*, edited by M. E. H. Schouten (Mouton de Gruyter, Berlin, Germany), pp. 85–98.
Li, F., Menon, A., and Allen, J. B. (**2010**). "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," J. Acoust. Soc. Am. **127**(4), 2599–2610.
Li, F., Trevino, A., Menon. A., and Allen, J. B. (**2012**). "A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise," J. Acoust. Soc. Am. **132**(4), 2663–2675.
Lobdell, B. E., and Allen, J. B. (**2007**). "A model of the VU (volume-unit) meter, with speech applications," J. Acoust. Soc. Am. **121**(1), 279–285.
Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. (**2014**). "Mechanisms of noise robust representation of speech in primary auditory cortex," Proc. Natl. Acad. Sci. **111**(18), 6792–6797.
Messing, D. P., Delhorne, L., Bruckert, E., Braida, L. D., and Ghitza, O. (**2009**). "A non-linear efferent-inspired model of the auditory system; matching human confusions in stationary noise," Speech Commun. **51**, 668–683.
Meyer, B. T., Brand, T., and Kollmeier, B. (**2011**). "Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes," J. Acoust. Soc. Am. **129**(1), 388–403.
Miller, G. A., and Licklider, J. C. R. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**(2), 167–173.
Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**(2), 338–352.
Moore, B. C. J. (**2003**). "Temporal integration and context effects in hearing," J. Phonetics **31**, 563–574.
Nielsen, J. B., and Dau, T. (**2009**). "Development of a Danish speech intelligibility test," Int. J. Audiol. **48**(10), 729–741.
Nielsen, J. B., and Dau, T. (**2011**). "The Danish Hearing in noise test," Int. J. Audiol. **50**(3), 202–208.
Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**(2), 1085–1099.
Payton, K. L., and Braida, L. D. (**1999**). "A method to determine the speech transmission index from speech waveforms," J. Acoust. Soc. Am. **106**(6), 3637–3648.
Phatak, S. A., and Allen, J. B. (**2007**). "Consonant and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am. **121**(4), 2312–2326.
Phatak, S. A., Lovitt, A., and Allen, J. B. (**2008**). "Consonant confusions in white noise," J. Acoust. Soc. Am. **124**(2), 1220–1233.
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**(6), 3988–3997.
Sakoe, H., and Chiba, S. (**1978**). "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoust. Speech Signal Process. ASSP-**26**, 43–49.
Singh, R., and Allen, J. B. (**2012**). "The influence of stop consonants' perceptual features on the articulation index model," J. Acoust. Soc. Am. **131**(4), 3051–3068.
Sroka, J. J., and Braida, L. D. (**2005**). "Human and machine consonant recognition," Speech Commun. **45**, 401–423.
Toscano, J. C., and Allen, J. B. (**2014**). "Across- and within-consonant errors for isolated syllables in noise," J. Speech Lang. Hear. Res. **57**, 2293–2307.
Tóth, M. A., García Lecumberri, M. L., Tang, Y., and Cooke, M. (**2015**). "A corpus of noise-induced word misperceptions for Spanish," J. Acoust. Soc. Am. **137**(2), EL184–EL189.
Wagener, K., Jovassen, J. L., and Ardenkjær, R. (**2003**). "Design, optimization and evaluation of a danish sentence test in noise," Int. J. Audiol. **42**, 10–17.
Wang, M. D., and Bilger, R. C. (**1973**). "Consonant confusions in noise: A study of perceptual features," J. Acoust. Soc. Am. **54**(5), 1248–1266.
Warren, R. M. (**1970**). "Perceptual restoration of missing speech sounds," Science **167**, 392–393.
Zaar, J., and Dau, T. (**2015**). "Sources of variability in consonant perception of normal-hearing listeners," J. Acoust. Soc. Am. **138**(3), 1253–1267.