Auditory correlates of stimulus-induced variability in consonant perception

Johannes Zaar[1] and Torsten Dau[2]
[1,2]*Hearing Systems, Department of Electrical Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark*
e-mail: [1]jzaar@elektro.dtu.dk, [2]tdau@elektro.dtu.dk

Introduction

Speech perception is often studied from a *macroscopic* perspective, i.e., in terms of the percentage of correctly identified meaningful words or sentences presented in a given condition (e.g. in additive noise). In a macroscopic speech intelligibility test, the acoustic information contained in the stimuli represents only one of several cues utilized by the listeners. If the acoustic information is degraded, lexical, semantic, and/or syntactic information (depending on the experimental design) is exploited by the listeners to recognize a given word. To investigate solely the relation between the acoustic properties of the stimulus and the resulting speech percept in a more controlled manner, a *microscopic* perspective may be taken by measuring consonant perception. Here, nonsense syllables like consonant-vowel combinations (CVs) are typically presented to listeners in additive steady-state noise. The responses to each speech stimulus are then analyzed both in terms of consonant *recognition* and consonant *confusions*.

Several concepts for modeling macroscopic speech intelligibility have been proposed. Some of these shall be mentioned here as representatives of the two main concepts in speech perception modeling: The traditional Articulation Index (AI) [1] is based on the signal-to-noise ratio (SNR) and audibility at the output of an auditory inspired filterbank and represents the *audibility*-based approach. The more recent Extended Speech Intelligibility Index (ESII) [2] is essentially a short-term version of the AI. The speech-based Envelope Power Spectrum Model (sEPSM) [3], on the other hand, considers speech intelligibility to be proportional to the SNR in the modulation domain ($\mathrm{SNR}_{\mathrm{env}}$) and represents the *modulation-masking* based approach.

A few recent studies have attempted to relate microscopic consonant perception data to the above mentioned classes of models. Li et *al.* [4] related consonant recognition data to the so-called AI Gram, a short-term representation of the AI that is conceptually comparable to the ESII. Jürgens and Brand [5] used an elaborate auditory model with a 4-channel modulation filterbank to predict consonant recognition and confusions. A template-matching back end based on a dynamic time warping (DTW) algorithm [6] was used and different back end configurations were tested. The model was shown to account well for consonant recognition while the confusion predictions were less successful. A similar concept was applied in a study by Zaar et *al.* [7], which compared the predictive power of different front ends using a fixed DTW-based back end configuration. It was shown that a modulation-domain front end yielded more accurate consonant recognition predictions than an audibility-based front end. However, the confusion predictions obtained with both front ends were found to be unsatisfactory.

The present study took an alternative approach to test the suitability of different auditory models for consonant perception modeling. First, consonant perception data were obtained with Danish normal-hearing listeners. The data were analyzed with respect to four different potential sources of stimulus-induced variability using a measure of the perceptual distance between responses. In particular, the perceptual distances induced by the acoustical differences (i) across CVs (i.e., across stimuli of *different* phonetic identity), (ii) across talkers, (iii) within talkers (both for stimuli of the *same* phonetic identity), and (iv) across masking-noise tokens (mixed with identical speech tokens) were calculated. Then, the corresponding stimuli were fed through different audibility- and modulation-based auditory models and the distances between the obtained internal representations were calculated using DTW. Finally, the perceptual distances were compared to the corresponding modeled distances. The suitability of the different modeling approaches is discussed.

Sources of perceptual variability

Experiment 1: Speech variability. 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. Six recordings of each CV (three spoken by a male, three spoken by a female talker) were taken from the Danish nonsense syllable speech material collected by Christiansen and Henrichsen [8], cut, and faded in and out manually. The levels were equalized using VUSOFT, a software implementation of an analog VU-meter [9], which effectively equalizes the vowel levels and thus ensures realistic relations between the levels of the consonants. One particular white masking noise waveform with a duration of one second was generated for each speech token in each SNR condition and faded in and out using raised cosine ramps with a duration of 50 ms. SNR conditions of 12, 6, 0, -6, -12, and -15 dB were created by fixing the noise level and adjusting the overall root-mean-square level of the speech tokens according to the desired SNR. 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.

Experiment 2: Noise variability. For each type of CV from experiment 1, only one recording spoken by the male talker was used. The equalization was performed as described above. Three masking-noise conditions (frozen noise A, frozen noise B, and random noise) were considered. For each speech token, one particular white noise waveform with a duration of one second was generated and labeled "frozen noise A"; the same noise token was then circularly shifted in time by 100 ms to obtain "frozen noise B". The noise tokens were faded in and out using raised cosine ramps with a duration of 50 ms. The noise waveforms for the random noise condition (added to prevent noise learning) were newly generated for each presentation and faded in and out in the same manner during the experimental procedure. The noisy speech tokens for the SNR conditions (12, 6, 0, -6, -12, and -15 dB) were created as in experiment 1.

Procedure. Eight normal-hearing native Danish listeners with an average age of 25 years participated in the experiment. Listeners were presented with the stimuli in experimental blocks ordered according to SNR in descending order. Each block included a short training run. The order of presentation within one experimental block was randomized. In experiment 1, each stimulus (each noisy speech token at each SNR) was presented three times to each listener. In experiment 2, each stimulus (each speech token in each masking noise condition at each SNR) was presented five times to each listener. Listeners were seated in a sound attenuating listening booth in front of a computer display and listened to the stimuli monaurally through equalized Sennheiser HD580 headphones. The stimuli were played as ".wav" files (44.1 kHz, 16 bits). The sound pressure level of the noise was set to 60 dB, while the overall stimulus level differed depending on the level of the speech (i.e., on the SNR). After each stimulus presentation, listeners had to choose one of the response alternatives displayed as 15 buttons labeled /b, d, f, g, h, j, k, l, m, n, p, s, ʃ, t, v/ and one button labeled "I don't know" on a graphical user interface (GUI).

Perceptual distance measure. For each stimulus and listener, the responses obtained in the experiments were converted to proportions of responses by distributing any "I don't know" response evenly across the 15 other response alternatives and dividing the frequencies of reponses by the number of stimulus presentations. The perceptual distance between two response vectors $\mathbf{r_1}$ and $\mathbf{r_2}$ was defined as the normalized angular distance between them:

$$\mathrm{D}(\mathbf{r_1},\mathbf{r_2}) = \arccos\left(\frac{\langle\mathbf{r_1},\mathbf{r_2}\rangle}{||\mathbf{r_1}||\cdot||\mathbf{r_2}||}\right)\cdot\frac{100\%}{\pi/2}, \tag{1}$$

where $\mathbf{r_i} = [\mathrm{p_b}, \mathrm{p_f}, ..., \mathrm{p_v}]$ denotes the response vector obtained for stimulus i and $\mathrm{p_x}$ represents the proportion of response "x". The normalization term contains the maximum possible angle of $\pi/2$ and re-scales the result to a percentage.
The perceptual distance was calculated across four different factors: (i) across CVs (i.e., across stimuli of *different* phonetic identity), (ii) across talkers, (iii) within talkers (both for stimuli of the *same* phonetic identity), and (iv) across masking-noise tokens (mixed with identical speech tokens). The across-noise distance was calculated using the data obtained in experiment 2. All other distances were extracted from the data obtained in experiment 1. For each considered factor, the perceptual distance was calculated across all pairwise comparisons of response vectors representative of that factor. The calculation was performed for each SNR condition separately and the individual distance values were averaged across the considered response pairs and across listeners. As a result, the across-CV, across-talker, within-talker, and across-noise perceptual distances were obtained as a function of the SNR.

Results. Figure 1 shows the results. As expected, the largest perceptual distance was observed across CVs (black bars). While the across-CV distance was at ceiling for large SNRs (as correctly recognized stimuli resulted in orthogonal response vectors), it decreased with decreasing SNR (as listeners made more confusions). However, different speech tokens of the same phonetic identity also produced substantial perceptual distances as reflected in the across-talker (blue bars) and within-talker (green bars) cases. These distances were low for large SNRs (as correctly recognized stimuli resulted in similar response vectors) and increased towards lower SNRs (as listeners made more confusions). Even a time shift in the masking-noise waveforms mixed with the same speech token led to a measurable perceptual distance (red bars) that increased with decreasing SNR. The largest CV-specific distance was the across-talker distance (blue bars), followed by the within-talker distance (green bars); the smallest effect was found for the across-noise distance (red bars). This ranking remained almost constant across SNR.

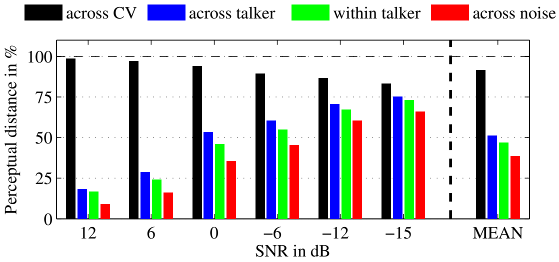


Figure 1: Perceptual distances across CVs (black), across talkers (blue), within talkers (green), and across noise (red) as a function of the SNR. The rightmost cluster shows the average across SNR.

Sources of variability in a model framework

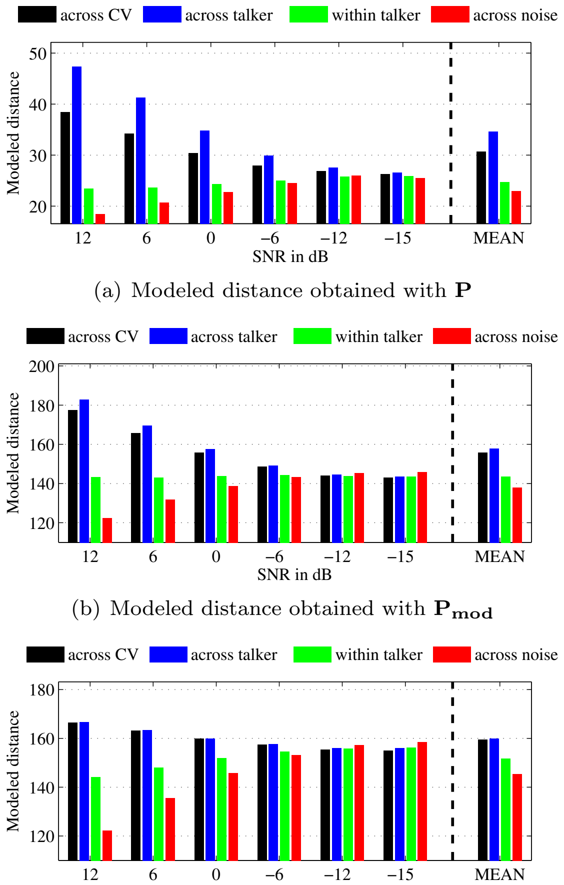
Subband power P. The subband power **P** was calculated using 22 gammatone filters with equivalent rectangular bandwidths. The gammatone filter center frequencies f_c were spaced on a third-octave grid, covering a range from 63 Hz to 8 kHz. The Hilbert envelope of the temporal output of each filter was extracted and low-pass filtered using a first-order Butterworth filter with a cut-off frequency of 150 Hz. The subband envelopes were downsampled to a sampling frequency of 1050 Hz and the power of the subband envelopes was calculated and converted to dB. $\mathbf{P}(t, fc)$ is a function of time t and gammatone filter center frequency f_c.

Modulation power $\mathbf{P_{mod}}$. The modulation power $\mathbf{P_{mod}}$ was obtained using the same subband decomposition, envelope extraction, lowpass filtering, and downsampling as described above. Each subband envelope was then passed through a modulation filterbank consisting of seven second-order bandpass filters in parallel with one lowpass filter. The bandpass filter center frequencies were octave spaced between 4 Hz and 256 Hz. The modulation lowpass filter was of third order with a cut-off frequency of 2 Hz. The power at the output of each modulation filter was calculated in dB. $\mathbf{P_{mod}}(t, f_c, f_m)$ is a function of time t, gammatone filter center frequency f_c, and modulation frequency f_m.

AC-coupled modulation power $\mathbf{P^{ac}_{mod}}$. The ac-coupled modulation power $\mathbf{P^{ac}_{mod}}$ was obtained in the same way as the modulation power $\mathbf{P_{mod}}$. However, the output of each modulation filter was in this case normalized by the long-term subband DC, which is consistent with the sEPSM by Jørgensen et *al.* [3]. $\mathbf{P^{ac}_{mod}}(t, f_c, f_m)$ is a function of time t, gammatone filter center frequency f_c, and modulation frequency f_m.

Modeled distance calculation. The experimental stimuli (excluding the noise-only portions at beginning and end) were fed through each of the models and the corresponding internal representations (IRs) were obtained. A standard dynamic time warping (DTW) algorithm was applied [6] to obtain the modeled distance between the IRs. It is based on a distance matrix $D(t_1, t_2)$ that contains the Euclidean distances between the IRs of two stimuli for all possible combinations of temporal samples. The DTW algorithm finds the path through this matrix that results in the minimum possible cumulative distance along its elements, using a dynamic programming scheme for efficiency. The Euclidean distance matrices were calculated as $\mathrm{D}_{\mathbf{P}}(t_1, t_2) = \sqrt{\sum_{f_c} [\mathbf{P_1}(t_1, f_c) - \mathbf{P_2}(t_2, f_c)]^2}$ for $\mathbf{P}$ and as $\mathrm{D}_{\mathbf{P_{mod}}}(t_1, t_2) = \sqrt{\sum_{f_c} \sum_{f_m} [\mathbf{P_{mod,1}}(t_1, f_c, f_m) - \mathbf{P_{mod,2}}(t_2, f_c, f_m)]^2}$ for $\mathbf{P_{mod}}$. $\mathrm{D}_{\mathbf{P^{ac}_{mod}}}$ was calculated similarly to $\mathrm{D}_{\mathbf{P_{mod}}}$. The modeled distance was defined as the cumulative distance between two IRs obtained using DTW. To minimize the influence of differences in duration, the cumulative distance was normalized by the length of the alignment path. The modeled distances were calculated across all pairwise comparisons of stimuli that had also been considered for the perceptual distance calculation. The calculation was performed for each SNR condition separately and the individual distance values were averaged across the considered stimulus pairs. As a result, the across-CV, across-talker, within-talker, and across-noise modeled distances were obtained as a function of the SNR.

Results. Figures 2(a), 2(b), and 2(c) show the modeled distances obtained with the three different models. It can be seen that the trends that were observed in Figure 1 for the across-CV (black bars), within-talker (green bars), and across-noise (red bars) perceptual distances were well captured by all three models. However, the modeled across-talker distances (blue bars in Figure 2) strongly deviated from the perceptual across-talker distances (blue bars in Figure 1): while the perceptual data showed an increasing distance with decreasing SNR, all simulations showed the reversed trend. Furthermore, the across-talker distance values were strongly overestimated by all models. This overestimation was most pronounced for $\mathbf{P}$, less pronounced for $\mathbf{P_{mod}}$, and least pronounced for $\mathbf{P^{ac}_{mod}}$, as reflected in the blue bars in Figures 2(a), 2(b), and 2(c), respectively.



(a) Modeled distance obtained with $\mathbf{P}$

(b) Modeled distance obtained with $\mathbf{P_{mod}}$

(c) Modeled distance obtained with $\mathbf{P^{ac}_{mod}}$

Figure 2: Modeled distances across CVs (black), across talkers (blue), within talkers (green), and across noise (red) as a function of the SNR. The rightmost cluster shows the average across SNR.

Model parameter evaluation

To investigate the influence of the considered envelope bandwidth, the modeled distances were calculated multiple times for each model with different model parameters. In particular, the distances were calculated (i) for $\mathbf{P}$ using 8 different envelope lowpass filter cut-off frequencies (2, 4, 8, 16, 32, 64, 128, 256 Hz) and (ii) for $\mathbf{P_{mod}}$ and $\mathbf{P^{ac}_{mod}}$ using 8 different modulation filterbank configurations (only first, first two, first three, etc. filters of the modulation filterbank). Pearson‘s correlation coefficient r was calculated between the obtained perceptual distance pattern depicted in Figure 1 and the respective modeled distance patterns. This reflects qualitative similarity of the distance distributions across SNR and across the considered distance types.

Figure 3 shows the performance of the different models in terms of Pearson‘s r as a function of the lowpass filter cut-off frequency (in case of $\mathbf{P}$) and the center frequency of the highest modulation filter considered in the modulation filterbank (in case of $\mathbf{P_{mod}}$ and $\mathbf{P}^{ac}_{\mathbf{mod}}$), respectively. The lowest similarity between the perceptual and the modeled distances was observed for $\mathbf{P}$ (squares), which increased towards larger lowpass filter cut-off frequencies. $\mathbf{P_{mod}}$ (diamonds) yielded a better match to the perceptual distance, especially when using the modulation filters up to 8 Hz; higher-frequency modulation filters slightly worsened the performance. For $\mathbf{P}^{ac}_{\mathbf{mod}}$ (circles), the match with the perceptual distance was found to be by far the best of all models, particularly when using only the low-frequency modulation filters.

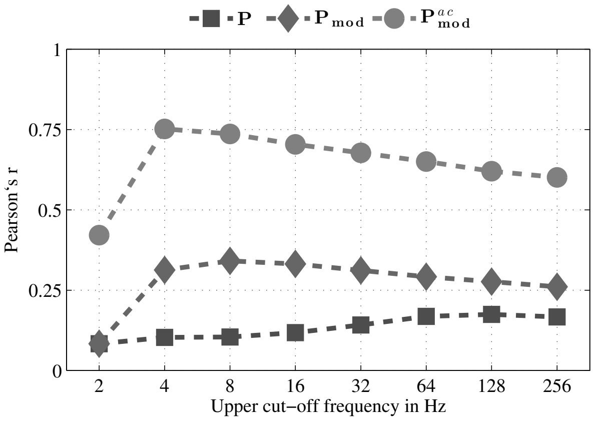


Figure 3: Pearson’s r calculated between perceptual distance pattern and modeled distance patterns as a function of permitted envelope bandwidth for subband power $\mathbf{P}$ (squares), modulation power $\mathbf{P_{mod}}$ (diamonds), and ac-coupled modulation power $\mathbf{P}^{ac}_{\mathbf{mod}}$ (circles).

Summary and discussion

Four sources of stimulus-induced variability in consonant perception were considered: across-CV, across-talker, within-talker, and across-noise variability. The influence of these sources of variability on the responses of normal-hearing listeners was quantified using a perceptual distance measure. The distances between the corresponding stimuli, as interpreted by three different auditory models, were obtained and compared to the perceptual distances. While the across-CV, within-talker, and across-noise distances found in the perceptual data were well represented by all models, the models strongly overestimated the across-talker distance for large SNRs. A closer inspection suggested that the models overestimated the contribution of long-term spectral differences between talkers of different gender. Such talker-specific differences represent a challenge in automatic speech recognition, where explicit speaker normalization techniques are typically applied to mitigate to effect [10]. In the present study, the overestimation was least pronounced using a modulation filterbank analysis followed by a normalization of the filter outputs by the long-term subband DC (as in the sEPSM by Jørgensen et *al.* [3]). This seems plausible as only relative changes in the subband envelopes are seen by such a model.

Overall, the audibility-based model showed a weak correlation with the perceptual domain. Introducing a modulation filterbank yielded a substantial increase in the correlation of model domain and perceptual domain. Further introducing normalization of the modulation filterbank outputs by the corresponding long-term subband DC resulted in the by far best observed match.

Acknowledgements

This research was supported by the FP7 Marie Curie Initial Training Network INSPIRE.

References

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

[2] Rhebergen, K. and Versfeld, N.: 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 (2005), 2181-2192.

[3] Jørgensen, S., Ewert, S., Dau, T.: A multi-resolution envelope-power based model for speech intelligibility. J. Acoust. Soc. Am. 134 (2013) 436-446.

[4] Li, F., Menon, A., Allen, J.: A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. J. Acoust. Soc. Am. 127 (2010) 2599-2610.

[5] Jürgens, T. and Brand, T.: Microscopic prediction of speech recognition for listeners with normal hearing using an auditory model. J. Acoust. Soc. Am. 126 (2009) 235-2648.

[6] Sakoe, H. and Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust., Speech, Signal Process. ASSP-26 (1978) 43-49.

[7] Zaar, J., Jørgensen, S., and Dau, T.: Modeling consonant perception in normal-hearing listeners. Proc. Forum Acusticum (2014).

[8] Christiansen, T. and Henrichsen, P.: Objective Evaluation of Consonant-Vowel pairs produced by native speakers of Danish. Proc. Forum Acusticum 2011, 67-72.

[9] Lobdell, B., Allen, J.: A model of the VU (volume-unit) meter, with speech applications. J. Acoust. Soc. Am. 121 (2007) 279-285.

[10] Lee, Li and Rose, R.: A frequency warping approach to speaker normalization. Speech and Audio Processing, IEEE Transactions on, (1998) 49-60.