Sources of Variability in Consonant Perception and Implications for Speech Perception Modeling

Johannes Zaar and Torsten Dau

Abstract The present study investigated the influence of various sources of response variability in consonant perception. A distinction was made between source-induced variability and receiver-related variability. The former refers to perceptual differences induced by differences in the speech tokens and/or the masking noise tokens; the latter describes perceptual differences caused by within- and across-listener uncertainty. Consonant-vowel combinations (CVs) were presented to normal-hearing listeners in white noise at six different signal-to-noise ratios. The obtained responses were analyzed with respect to the considered sources of variability using a measure of the perceptual distance between responses. The largest effect was found across different CVs. For stimuli of the same phonetic identity, the speech-induced variability across and within talkers and the across-listener variability were substantial and of similar magnitude. Even time-shifts in the waveforms of white masking noise produced a significant effect, which was well above the within-listener variability (the smallest effect). Two auditory-inspired models in combination with a template-matching back end were considered to predict the perceptual data. In particular, an energy-based and a modulation-based approach were compared. The suitability of the two models was evaluated with respect to the source-induced perceptual distance and in terms of consonant recognition rates and consonant confusions. Both models captured the source-induced perceptual distance remarkably well. However, the modulation-based approach showed a better agreement with the data in terms of consonant recognition and confusions. The results indicate that low-frequency modulations up to 16 Hz play a crucial role in consonant perception.

J. Zaar (✉)
Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, Ørsteds Plads Building 352, room 108, 2800 Kongens Lyngby, Denmark
e-mail: jzaar@elektro.dtu.dk

T. Dau
Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, Ørsteds Plads Building 352, room 120, 2800 Kongens Lyngby, Denmark
e-mail: tdau@elektro.dtu.dk

P. van Dijk et al. (eds.), *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing,* Advances in Experimental Medicine and Biology 894,
DOI 10.1007/978-3-319-25474-6_46

Keywords Talker effects · Noise effects · Listener effects · Internal noise · Auditory modeling · Perceptual distance · Envelope domain · Microscopic · Perceptual distance · Modulation filters

1 Introduction

Speech perception is often studied from a *macroscopic* perspective, i.e., using meaningful long-term speech stimuli (e.g., in additive noise). To solely investigate the relation between the acoustic properties of the stimulus and the resulting speech percept (excluding lexical, semantic, and syntactic effects), an alternative is to take a *microscopic* perspective by investigating the perception of smaller units of speech such as consonants. Miller and Nicely (1955) measured the perception of consonant-vowel combinations (CVs, e.g.,/ba/,/ta/) in white noise and different band-pass filtering conditions and observed distinct consonant confusions. Wang and Bilger (1973) demonstrated that consonant perception also depends on the vowel context. Phatak and Allen (2007) measured consonant perception in speech-weighted noise and demonstrated noise-type induced perceptual differences to the Miller and Nicely (1955) data. In following studies, perceptual differences across different speech tokens of the same phonetic identity came more into focus (e.g., Phatak et al. 2008).

A few studies have attempted to simulate consonant perception. Li et al. (2010) successfully related consonant recognition data to the so-called AI Gram, which is related to the energy-based Articulation Index (ANSI 1969). Gallun and Souza (2008) considered noise-vocoded VCVs and demonstrated that the correlation of long-term modulation power representations was a strong predictor of consonant confusions. Jürgens and Brand (2009) used an auditory model with a modulation-frequency selective preprocessing stage in combination with a template-matching back end. The model showed convincing recognition predictions while the confusion predictions were inconclusive.

Motivated by the increasing evidence for a major variability in consonant perception that cannot be accounted for by the phonetic identity of the stimuli, the present study attempted to quantify some of the sources of variability that influence consonant perception. It was distinguished between *source-induced variability* and *receiver-related variability*. The former was subdivided into speech- and noise-induced variability; the latter was subdivided into across- and within-listener variability. Consonant perception data were collected and analyzed with respect to the considered sources of variability using a measure of the perceptual distance between responses. Predictions of the data were obtained using an energy- and a modulation-based model in combination with a template-matching back end. The model predictions were compared to the data (i) in terms of how well they reflected the source-induced variability measured in listeners and (ii) in terms of the agreement between perceptual and predicted consonant recognition and confusions.

2 Methods

2.1 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/ (as in "feed") 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 (2011). Six SNR conditions (12, 6, 0, −6, −12, and −15 dB) were created by fixing the noise sound pressure level (SPL) to 60 dB and adjusting the speech SPL. One particular white masking noise waveform with a duration of 1 s was generated for each speech token in each SNR condition and mixed with it such that the speech token onset was temporally positioned 400 ms after the noise onset.

2.2 Experiment 2: Noise Variability

Only one male-talker speech token of each CV was used. 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 1 s 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 waveforms for the random noise condition (added to prevent noise learning) were randomly generated during the experimental procedure. The noisy speech tokens were created as described in Sect. 2.1.

2.3 Experimental Procedure

Two different groups of eight normal-hearing native Danish listeners participated in the two experiments (average age: 26 and 24 years, respectively). The stimuli were monaurally presented to the listeners via headphones in experimental blocks ordered according to the SNR in descending order. An additional quiet condition (clean speech at 60 dB SPL) preceded the SNR conditions (noise level at 60 dB SPL). Each block started with a short training run. The order of presentation within each experimental block was randomized. In experiment 1, each stimulus was presented three times to each listener. In experiment 2, each stimulus was presented five times to each listener. 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, Sj, t, v" and one button labeled "I don't know" on a graphical user interface (the Danish "Sj" corresponds to /ʃ/). Experiment 2 was repeated with four of the originally eight listeners to obtain test-retest data.

2.4 Data Analysis

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 occurrences of responses by the number of stimulus presentations. The response of a given listener obtained with a given stimulus was thus calculated as a vector $r=[\mathrm{p_b}, \mathrm{p_d}, \ldots, \mathrm{p_v}]$, where $\mathrm{p_x}$ denotes the proportion of response “x”. The perceptual distance between two response vectors r_1 and r_2 was defined as the normalized angular distance between them:

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

The perceptual distance was calculated across six different factors: (i) across CVs, (ii) across talkers, (iii) within talkers, (iv) across masking-noise tokens, (v) across listeners, and (vi) within listeners. Apart from the across-CV factor, only responses obtained with stimuli of the same phonetic identity were compared. For each considered factor, the perceptual distance was calculated across all pairwise comparisons of response vectors representative of that factor. The calculations for all factors but (v) were performed for each listener and each SNR condition separately. The calculations for (v) were performed for each SNR condition separately, comparing responses across listeners. The individual distance values were then averaged across the considered response pairs and (where applicable) across listeners. As a result, the respective perceptual distances were obtained as a function of SNR.

3 Experimental Results

Figure 1 shows examples of perceptual *across-talker variability* and perceptual *across-noise variability* in terms of across-listener average confusion patterns: /pi/ spoken by talker A (panel a) was more recognizable (and less confusable) than /pi/ spoken by talker B (panel b); the perception of a given speech token /gi/ was very differently affected by a specific white masking-noise waveform “A” (panel c) than by a time-shifted version of that waveform (“B”, panel d).

Figure 2 shows the perceptual distances derived from the experimental data for all SNR conditions (see Sect. 2.4). On the left, the across-SNR average is shown. The largest perceptual distance of 91 % was found across CVs (black bar). Regarding the source-induced perceptual distances across stimuli of the same phonetic identity, the largest perceptual distance of 51 % was obtained across talkers (blue bar), followed by the within-talker factor (47 %, green bar). A temporal shift in the masking-noise waveform induced a perceptual distance of 39 % (red bar). Regarding the receiver-related effects, a substantial perceptual distance of 46 % across

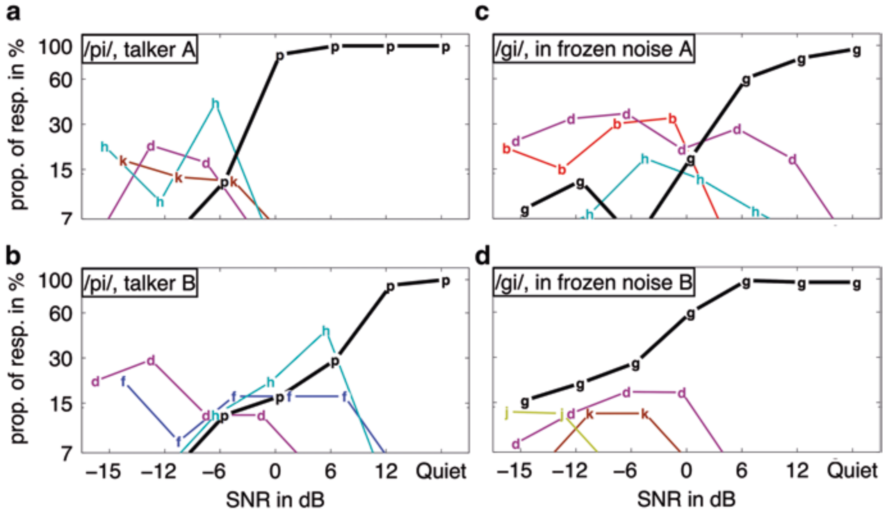

Fig. 1 Across-listener average example confusion patterns (CPs). *Left*: CPs obtained with two different speech tokens /pi/ spoken by male talker A (*top*) and female talker B (*bottom*). *Right*: CPs obtained with one specific speech token /gi/ mixed with frozen noise A (*top*) and frozen noise B (*bottom*)

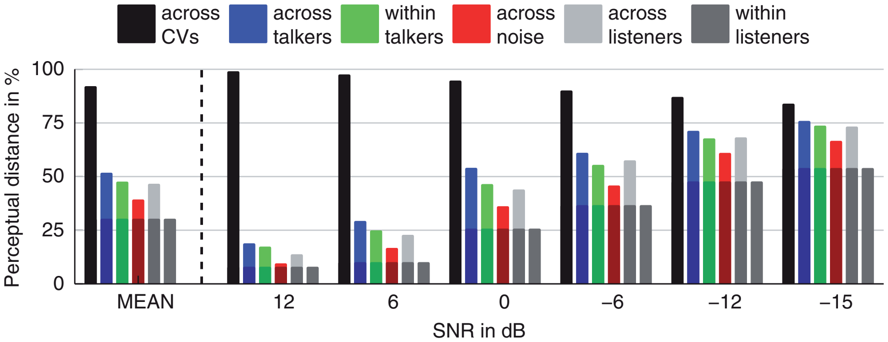

Fig. 2 Mean (*left*) and SNR-specific perceptual distances across CVs, across talkers, within talkers, across noise, across listeners, and within listeners. The *shaded* areas represent values below the within-listener distance, i.e., below the internal-noise baseline

listeners was found for physically identical stimuli (light gray bar). In contrast, the relatively low perceptual distance of 30 % within listeners (test vs. retest, dark gray bar) indicated that the individual listeners were able to reproduce their responses fairly reliably. Pairwise t-tests across all combinations of conditions (excluding the across-CV condition) demonstrated that all conditions were significantly different from each other ($p < 0.05$) except for the across-talker (blue), within-talker (green), and across-listener (light gray) conditions.

Regarding the trends across SNR in Fig. 2, it can be seen that the across-CV distance (black bars) was at ceiling for large SNRs and decreased with decreasing SNR, as listeners made more speech-token specific confusions. All other perceptual distance types showed low values for large SNRs and increased with decreasing SNR due to stimulus-specific confusions and listener uncertainty. The within-listener distance ("internal noise") represented the baseline and strongly increased with decreasing SNR as the task became more challenging.

4 Modeling

4.1 Model Components

The subband power P, in dB, was calculated using 22 fourth-order gammatone filters with equivalent rectangular bandwidths. The center frequencies were spaced on a third-octave grid, covering a range from 63 Hz to 8 kHz. The Hilbert envelope of each filter output was extracted and low-pass filtered using a first-order Butterworth filter with a cut-off frequency of 150 Hz. The envelopes were down-sampled to a sampling rate of 1050 Hz.

The modulation power P_{mod}, in dB, was obtained using the subband envelope extraction described above, followed by a modulation filterbank consisting of 3 second-order band-pass filters (center frequencies of 4, 8, and 16 Hz) in parallel with one third-order low-pass filter (cut-off frequency of 2 Hz).

A template-matching procedure was applied to predict the responses obtained in experiment 1. Two talker-specific template sets were considered, consisting of all speech tokens from each talker (i.e., three templates for each CV). The templates were mixed with random white noise at the test-signal SNR and compared to the experimental stimuli. The distances between the models' internal representations of the test signals and the templates were obtained using a standard dynamic time warping (DTW) algorithm (Sakoe and Chiba 1978). The template-matching procedure was conducted nine times with newly generated random noise for the templates. The "correct" template always contained the same speech token as the test signal, while the masking noise differed. In each run, the template showing the smallest distance to the test signal was selected. The modeled responses were converted to proportions of responses. The responses obtained in experiment 2 were predicted similarly, considering only the 15 speech tokens used in experiment 2 as templates.

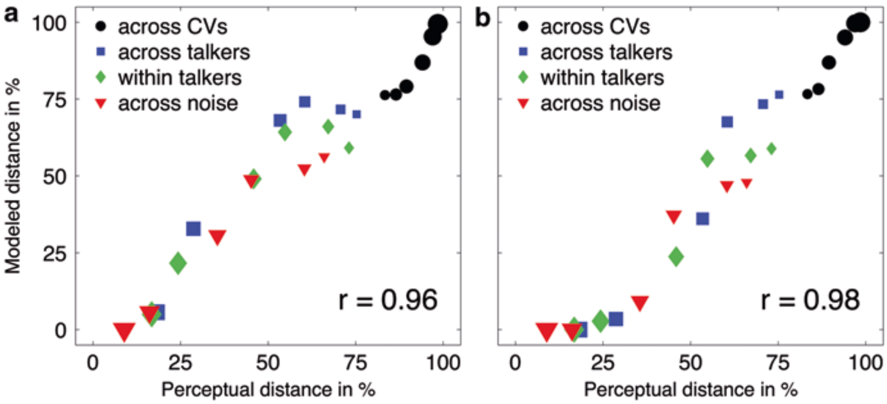

Fig. 3 Source-induced perceptual distances (from Fig. 2) plotted versus corresponding modeled distances obtained using P (*left*) and P_{mod} (*right*). The symbols and colors represent the different distance types, the size of the symbols is proportional to the SNR

5 Simulation Results

5.1 Sources of Variability

In accordance with the procedure described in Sect. 2.4, the across-CV, across-talker, within-talker, and across-noise modeled distances were obtained as a function of the SNR from the predicted responses. Figure 3 shows scatter plots of the perceptual distance versus the modeled distances obtained using P (panel a) and P_{mod} (panel b). It can be observed that the perceptual distances were remarkably well-predicted using P as well as P_{mod}, with a Pearson's *r* of 0.96 and 0.98, respectively.

5.2 Consonant Recognition and Confusions

The grand average consonant recognition rates obtained in experiment 1 showed a speech reception threshold (SRT) of −3 dB. The predicted SRTs were overestimated by 2.8 dB using P and by only 0.4 dB using P_{mod}. The token-specific SRTs showed a large spread across speech tokens, which was smaller in both model predictions. However, P_{mod} captured the relative ranking of token-specific SRTs considerably better than P (Spearman's *r* of 0.4 and 0.04, respectively).

The across-listener average data obtained in experiment 1 were averaged across different speech tokens of the same phonetic identity and across the four lowest SNRs (as most confusions occur for low SNRs). The resulting confusion matrices (CMs) are shown as filled gray circles in both panels of Fig. 4. The model predictions are plotted on top as open black circles. For P (panel a), an underestimation of the recognition for many consonants was observed, indicated by the mismatch

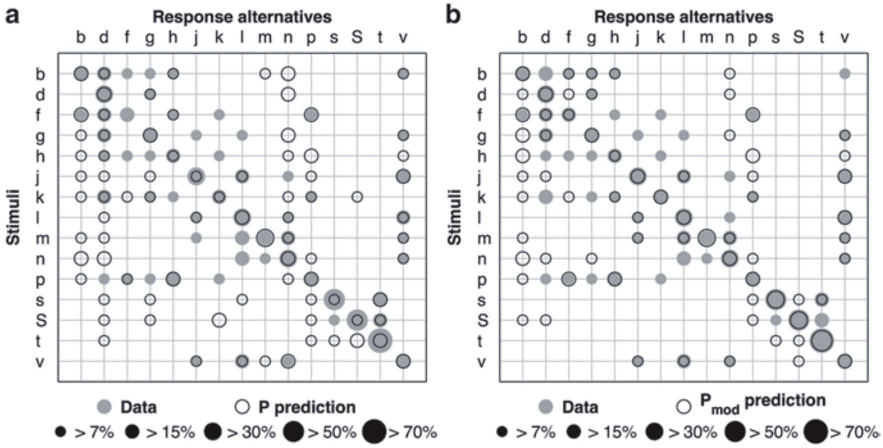

Fig. 4 Confusion matrices obtained in experiment 1, averaged across listeners, speech tokens of the same phonetic identity, and across SNRs of 0, −6, −12, and −15 dB. The perceptual data are shown as filled gray circles in both panels. The model predictions obtained with P (left) and P_{mod} (right) are represented as open *black circles*

of the on-diagonal circles. For P_{mod} (panel b), a good consonant-specific recognition rate match was found. Both models hit most of the confusions, reflected in the proportion of gray off-diagonal circles matched with a black circle. However, there were also many “false alarms”, particularly in the P-based predictions (panel a).

6 Summary and Discussion

The investigation of different sources of variability in Sect. 3 indicated that any considered difference in the stimuli produced a measurable effect. The observed perceptual variability across talkers is well established in the related literature (e.g., Phatak et al. 2008); the equally large variability within talkers had not yet been demonstrated. Most remarkably, even a 100-ms time shift in the white masking-noise waveform induced significant perceptual variability, indicating that “steady-state” masking noise should not be considered steady over time in the context of consonant cues. On the receiver side, different NH listeners with identical language background showed substantial differences while individual listeners could fairly reliably reproduce their responses. Averaging consonant perception data (even across NH listeners) thus seems problematic.

The predictions obtained in Sect. 4 with the energy-based (P) and the modulation-based (P_{mod}) pre-processing stages both accounted for the trends in the perceptual data with respect to the considered stimulus-related sources of variability. Consonant recognition was strongly underestimated using P and well-predicted using P_{mod}. An inspection of confusion matrices suggested that both models correctly predicted most of the perceptual confusions, albeit with some “false alarms”. The overall larger predictive power obtained with P_{mod} indicates that slow envelope fluctuations up to 16 Hz are a good predictor for consonant-in-noise perception. This is consistent with the findings by Gallun and Souza (2008).

The perceptual data analysis has implications for the further model design. It was shown that the internal noise increased with decreasing SNR. This could be incorporated in the model using an SNR-dependent random process in the decision stage (instead of SNR-dependent templates). Furthermore, the model predicted responses of a hypothetical “average” NH listener, which is unrealistic given the considerable across-listener variability. It remains a challenge to include listener-specific differences in the model, as it is not clear whether these differences can be accounted for by slight sensitivity differences between the NH listeners, cognitive effects, individual biases, or any combination of these factors. Eventually, an extension of the model towards different types of hearing impairment might be useful to understand the link between individual impairment factors and microscopic speech intelligibility.

Acknowledgments We thank Søren Jørgensen for his contributions to the research underlying this study. This research was supported by the FP7 Marie Curie Initial Training Network INSPIRE.

References

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

Christiansen TU, Henrichsen PJ (2011). Objective evaluation of consonant-vowel pairs produced by native speakers of Danish. Proceedings of Forum Acusticum 2011.

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

Jürgens T, 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

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

Miller GA, Nicely PE (1955) An analysis of perceptual confusions among some English consonants. J Acoust Soc Am 27(2):338–352

Phatak SA, Allen JB (2007) Consonant and vowel confusions in speech-weighted noise. J Acoust Soc Am 121(4):2312–2326

Phatak SA, Lovitt A, Allen JB (2008) Consonant confusions in white noise. J Acoust Soc Am 124(2):1220–1233

Sakoe H, Chiba S (1978). Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans Acoust, Speech Signal Proc (ASSP) 26(1), 43–49

Wang MD, Bilger RC (1973) Consonant confusions in noise: a study of perceptual features. J Acoust Soc Am 54(5):1248–1266