Sources of variability in consonant perception of normal-hearing listeners

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

(Received 21 December 2014; revised 24 July 2015; accepted 24 July 2015; published online 1 September 2015)

Responses obtained in consonant perception experiments typically show a large variability across stimuli of the same phonetic identity. The present study investigated the influence of different potential sources of this response variability. It was distinguished between source-induced variability, referring to perceptual differences caused by acoustical differences in the speech tokens and/or the masking noise tokens, and receiver-related variability, referring to perceptual differences caused by within- and across-listener uncertainty. Consonant-vowel combinations consisting of 15 consonants followed by the vowel /i/ were spoken by two talkers and presented to eight normal-hearing listeners both in quiet and in white noise at six different signal-to-noise ratios. The obtained responses were analyzed with respect to the different sources of variability using a measure of the perceptual distance between responses. The speech-induced variability across and within talkers and the across-listener variability were substantial and of similar magnitude. The noise-induced variability, obtained with time-shifted realizations of the same random process, was smaller but significantly larger than the amount of within-listener variability, which represented the smallest effect. The results have implications for the design of consonant perception experiments and provide constraints for future models of consonant perception.
[http://dx.doi.org/10.1121/1.4928142]

[JFC] Pages: 1253–1267

I. INTRODUCTION

Speech intelligibility is often characterized in terms of the percentage of correctly identified meaningful words or sentences presented to the listener, either in quiet or in the presence of a noise masker or interfering talker(s). For instance, a common measure of speech intelligibility is the speech reception threshold (SRT), which reflects the speech-to-masker/interferer energy at which 50% of the presented speech items have been correctly identified. The SRT measure may be considered as reflecting a *macroscopic* view on speech perception. The term macroscopic is threefold in the sense that (i) long-term speech units are used, such as words or sentences, (ii) only speech recognition is considered while confusions of words are not investigated, and (iii) meaningful speech is used, typically consisting of common words in a syntactically correct sentence structure. In this type of experimental setting, listeners can exploit information obtained from various stages of speech processing. For instance, missing acoustic information can be extrapolated using lexical, semantic and/or syntactic information. Approaches for measuring macroscopic speech intelligibility range from presenting syntactically diverse meaningful sentences as in the "hearing in noise test" (e.g., Nilsson *et al.*, 1994; Nielsen and Dau, 2011) and the "conversational language understanding evaluation" (Nielsen and Dau, 2009) to using matrix sentence tests (e.g., Hagerman, 1982; Wagener *et al.*, 2003), where semantically unpredictable sentences are presented within a fixed syntactical structure. Therefore, macroscopic speech intelligibility tests differ in the semantic and syntactic predictability provided, while lexical effects play a considerable role in any of these tests.

Addressing speech intelligibility at a more fundamental level, many studies have focused on investigating the perception of smaller units of speech, such as syllables or phones (i.e., consonants and vowels). The perception of vowels has been shown to be more robust in the presence of steady-state noise than the perception of many consonants (Phatak and Allen, 2007). Therefore, the most "critical" or vulnerable phones in this context are consonants. Combinations of consonants and vowels (e.g., /ta/, /ba/, etc.) have typically been considered and the perceptual data have been analyzed in terms of consonant recognition (i.e., the percentage of correctly identified consonants) as well as in terms of consonant confusions. This type of approach may be considered as *microscopic* as it (i) uses short-term speech stimuli, (ii) analyzes recognition as well as confusions, and (iii) employs nonsense speech stimuli, thus excluding the contribution of effects related to lexicon, meaning, and syntax. The microscopic approach therefore allows for an analysis of the mapping from the acoustical stimulus to the associated phone percept by minimizing the biases induced by higher-level speech processing. This could be relevant, for example, when analyzing the effects of acoustical transmission channels (e.g., mobile phones), hearing impairment, and hearing-aid signal processing algorithms on the perception of the fundamental building blocks of speech. However, in order to fully exploit the microscopic approach it seems

[a)]Electronic mail: tdau@elektro.dtu.dk

crucial to understand the factors that contribute to consonant perception.

The first investigations of nonsense syllable perception were conducted by Fletcher and colleagues in the context of their pioneering research on telephone speech transmission quality at the Bell Laboratories between 1919 and 1945 (e.g., Fletcher and Galt, 1950; see also Allen, 1994). Nonsense consonant-vowel-consonant (CVC), consonant-vowel (CV), and vowel-consonant (VC) combinations were used to assess the amount of correctly transmitted articulation under conditions of noise and spectral filtering. These investigations resulted in the definition of the articulation index (AI), a technical measure to determine the quality of speech transmission channels (French and Steinberg, 1947). Although Fletcher and Galt (1950) did not directly address phonetic confusions, their work provided the basis for further research on nonsense speech perception.

Miller and Nicely (1955) conducted the first study that focused on perceptual confusions among consonants. CVs consisting of the 16 most common English consonants followed by the vowel /a/ (as in father) were spoken by five talkers and presented to four listeners. In a set of experimental conditions, white noise was added at various signal-to-noise ratios (SNRs) and different band-pass filters were applied to the speech. After each presentation, listeners had to indicate the consonant they had heard. The responses were pooled across listeners and displayed as confusion matrices (CMs). Several perceptual confusion groups of consonants (e.g., /p, t, k/) were observed and the data were investigated in terms of the information transmitted by different articulatory features (voicing, nasality, affrication, duration, and place of articulation). Wang and Bilger (1973) considered CVs and VCs consisting of 25 consonants and the vowels /a, i, u/ and applied a sequential information analysis in an attempt to derive an ideal set of relevant articulatory features. However, their results suggested that an articulatory feature-based analysis might be inappropriate to account for the data. Furthermore, Wang and Bilger (1973) found that the accompanying vowel had an influence on the consonant detection performance as well as the type of consonant-vowel combination (CV or VC), demonstrating that consonant perception does not solely depend on the consonant but also on the vowel context the consonant is embedded in.

In a related more recent study, Allen (2005) re-analyzed the Miller and Nicely (1955) data and related them to the AI. Allen proposed that confusion matrices should be analyzed in terms of perceptual events rather than in terms of articulatory features. He introduced the confusion pattern (CP) which, for a given speech stimulus, depicts the proportions of the different response alternatives as a function of the experimental conditions (e.g., SNRs). The CP was shown to be more appropriate than the confusion matrix for characterizing perceptual confusion groups and other trends in the data since it provides an overview of the data across experimental conditions. Phatak *et al.* (2008) reproduced the main results obtained in the Miller and Nicely (1955) study and demonstrated considerable noise-type specific perceptual differences (comparing white noise and speech-weighted noise). They also showed that different speech tokens of the same phonetic identity induced strong differences in consonant recognition and confusions. Li *et al.* (2010) and Li *et* al. (2012) developed a psychoacoustic method named "three-dimensional deep search" which was designed to identify the spectro-temporal cue regions of consonants based on experimental consonant recognition data obtained with noise masking, spectral filtering, and time truncation. As this kind of microscopic speech investigation relies heavily on the characteristics of the individual speech tokens, the perceptual differences across different speech tokens of the same phonetic identity came more into focus.

Consistent with the findings of Phatak *et al.* (2008), Singh and Allen (2012) demonstrated the occurrence of major within-consonant speech-token specific differences in the recognition of stop consonants. Toscano and Allen (2014) investigated across- and within-consonant recognition errors for CVs consisting of the sixteen consonants used by Miller and Nicely (1955) followed by four different vowels. Each of the CVs was spoken by fourteen different talkers and presented at six SNRs in speech-weighted noise. The results suggested that consonant recognition greatly varies across consonants as well as within consonants (i.e., across talkers and accompanying vowels). This implies talker-dependent effects, which have also been shown for spoken word recognition (e.g., Mullenix *et al.*, 1989) and represent a major challenge in automatic speech recognition (Benzeghiba *et al.*, 2007).

While the above studies provided major insights into consonant perception from various perspectives, it has remained unclear (i) to what extent the reported speech-token dependence of consonant perception is related to articulatory differences across talkers or to differences in the accompanying vowel, (ii) how articulatory differences across different utterances of a given talker affect consonant perception, and (iii) whether spectro-temporal details of the individual masking-noise waveform affect consonant perception. Furthermore, perceptual differences across and within individual listeners have not yet been addressed systematically, apart from individual studies considering hearing-impaired (HI) listeners (e.g., Phatak *et al.*, 2009, Trevino and Allen, 2013) or groups of listeners with different language background (e.g., Cutler *et al.*, 2004).

The present study was undertaken in an attempt to quantify the relative importance of some of the factors that influence consonant perception both in terms of stimulus-related ("source") and listener-related ("receiver") effects. Here, a distinction was made between *source-induced* variability and *receiver-related* variability. Source-induced variability refers to perceptual differences that arise due to variations in the acoustic properties of the stimulus and is subdivided into (i) *speech-induced* variability (perceptual differences arising from articulatory differences in speech tokens of the same phonetic identity, categorized as across-talker and within-talker variability) and (ii) *noise-induced* variability (perceptual differences arising from differences in the waveform of the masking noise). Receiver-related variability refers to the uncertainty/variation of the perceptual response due to encoding/resolution differences and limits in the listeners and is subdivided into (i) *across-listener* variability and

(ii) *within-listener* variability. Additional well-known sources of variability like the position and type of the accompanying vowel, as well as the long-term spectral characteristics of the noise (e.g., white vs speech-weighted) were not considered here.

Fifteen Danish consonants combined with the vowel /i/ as CVs were used in the present study, spoken by non-professional native Danish talkers and presented to NH native Danish listeners. Two experiments were conducted using white noise maskers at six SNRs. Experiment 1 investigated the effect of variations in the speech stimulus using several speech tokens for each CV presented in deterministic white noise maskers. Experiment 2 addressed the effect of noise variability using only a single speech token per CV and presenting it in different deterministic realizations of white noise maskers. The experimental data were analyzed with respect to source-induced variability using the data obtained in experiment 1 for analyzing the speech-induced variability, and using the data obtained in experiment 2 for analyzing the noise-induced variability. Furthermore, the data obtained in the two experiments were analyzed with respect to receiver-related variability by comparing the responses to physically identical stimuli across and within listeners. The analyses were performed by comparing example confusion patterns and confusion matrices. The entire set of the collected data was furthermore analyzed using a perceptual distance measure and the entropy of responses to quantify the contributions of the different considered sources of variability.

II. METHOD

A. Experiment 1: Effects of variations in the speech stimulus

1. Listeners

Eight native Danish listeners (one female, seven male) with audiometric thresholds of 20 dB hearing level (HL) or less at the measured frequencies between 125 Hz and 8 kHz participated in the experiment. The age of the listeners ranged from 19 years to 27 years, except for one listener who was 38 years old. The average age was 26 years. Listeners were paid for their participation in the experiment.

2. Stimuli

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 throughout this study. For experiment 1, six recordings of each CV were taken from the Danish nonsense syllable speech material collected by 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. A total of 90 speech tokens was used in the experiment (15 CVs × 3 speech tokens × 2 talkers). The individual speech tokens were cut and faded in and out manually. Their levels were equalized using VUSOFT, a software implementation of an analog VU-meter developed by Lobdell and Allen (2007), which was also used for level equalization in Phatak *et al.* (2008). The level equalization was performed such that all CVs showed the same VUSOFT peak value. This equalization strategy is based on the vowel levels, thus ensuring realistic relations between the levels of the individual consonants. Therefore, the vowel levels across the equalized CVs were similar while the consonant levels differed, much like in natural speech. After equalization, the reference speech level for the SNR calculation was defined as the overall root-mean-square level of all speech tokens.

For the masking noise generation, a "half-frozen" noise approach was taken in order to avoid a potential blur in the perceptual data that might arise from an effect of differences in the noise waveforms. Specifically, the noise waveform was fixed ("frozen") for a given speech token in a given SNR condition and the waveform of the presented mixture of speech and noise was thus exactly the same across (i) repeated presentations of a speech token in a given SNR condition and (ii) across different listeners. For each speech token and each SNR condition, one white Gaussian masking noise token with a duration of 1 s was generated 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 level of the speech tokens based on the reference speech level according to the desired SNR. 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). This fixed noise level approach was chosen instead of the commonly used fixed speech level approach in order to avoid extremely high noise levels at low SNRs, which can lead to annoyance and fatigue in listeners. 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. The clean speech at the respective levels, the noise tokens, and the mixture of speech and noise tokens were individually stored in ".wav" format at a sampling rate of 44.1 kHz with a resolution of 16 bits per sample.

3. Experimental design

The experiment was split into two sessions, one using the 45 male talker speech tokens and the other one using the 45 female talker speech tokens. The listeners performed the individual sessions on different days. One session lasted approximately 2.5 h including instruction, training, and breaks. Two control conditions with speech in quiet were defined, "Q60" and "Q45." Q60 was designed to evaluate whether the speech tokens were sufficiently identifiable in ideal listening conditions; the speech was presented in quiet at the same presentation level at which the fixed-level noise was presented in the SNR conditions [60 dB sound pressure level (SPL)]. Since the noise level was fixed and the speech level was adapted to generate the individual SNR conditions, Q45 was designed to investigate whether the speech tokens were still sufficiently intelligible in quiet at the lowest speech level occurring in the SNR conditions; the speech was therefore presented at 45 dB SPL, corresponding to the speech level in the −15 dB SNR condition.

The experimental sessions were split into eight consecutive blocks corresponding to the eight experimental conditions. In order to get the listeners accustomed to the task, the first condition was the "easy" control listening condition Q60, followed by the slightly more challenging control condition Q45. The third to eighth conditions were the six SNR conditions ranked from easy to difficult, i.e., with SNR = 12, 6, 0, −6, −12, and −15 dB. Each block consisted of a training run followed by the experiment run. In the training run, all 45 stimuli (depending on the condition speech tokens or speech tokens mixed with the predefined noise tokens) were presented once in random order to familiarize the listener with the respective condition. In the experiment run, each of the 45 stimuli was presented 3 times, resulting in a total of 135 presentations. The order of presentation was again randomized. One experimental block therefore comprised 180, a whole session 1440 stimulus presentations.

4. Procedure and apparatus

Listeners were seated in a sound attenuating listening booth in front of a computer and listened to the stimuli monaurally through Sennheiser HD580 headphones. For headphone equalization, a 256-tab finite impulse response filter designed to equalize the third-octave smoothed version of the headphone transfer function in the range between 40 Hz and 21 kHz was applied. The test software was run under MATLAB on a Windows-based PC. The stimuli were played at a sampling rate of 44.1 kHz. After each stimulus presentation, listeners had to choose one of the response alternatives displayed on a graphical user interface (GUI). The task was identical in training and experiment with no feedback provided. When in doubt, the listeners could repeat the sound playback up to two times using a "repeat" button included in the GUI. The response alternatives consisted of 15 buttons displaying the consonants in the corresponding Danish spelling (b, d, f, g, h, j, k, l, m, n, p, s, Sj, t, v) and one button labeled "I don't know." Listeners could respond to the stimulus using a computer mouse. After a decision was made, the next stimulus was played after a 500-ms pause. Prior to the experiment, the listeners were instructed to make use of the "repeat" button whenever they were uncertain about their percept and to use the "I don't know" button instead of guessing when they had only heard the vowel. The proportions of responses for each speech token and condition were calculated via division of the obtained occurrences of responses by the number of stimulus presentations. The "I don't know" responses were evenly distributed across all response alternatives. The conversion was performed both for the pooled responses across listeners and the individual listeners' responses. Three observations per stimulus and individual listener were obtained; the number of pooled observations per stimulus was thus 24 (3 observations × 8 listeners).

B. Experiment 2: Effects of variations in the noise

1. Listeners

Eight native Danish listeners (one female, seven male) with audiometric thresholds of 20 dB hearing level (HL) or less at the measured frequencies between 125 Hz and 8 kHz participated in the experiment. Four of these listeners had also participated in experiment 1. The age of the listeners ranged from 20 to 28 years, with a mean age of 24 years. Listeners were paid for their participation in the experiment. To obtain test-retest data, a subset of the original listener panel (four of the eight listeners) conducted a retest approximately one month after the first test.

2. Stimuli

Experiment 2 addressed perceptual differences induced by different "frozen-noise" masker waveforms. For each type of CV from experiment 1, only one recording was used, resulting in 15 speech tokens. The recordings were a subset of the speech material used in experiment 1, spoken by the male talker. The level of the speech tokens was equalized according to the VUSOFT peak value (cf. Sec. II A 2). 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 tokens were faded in and out using raised cosine ramps with a duration of 50 ms. The noise waveforms for the random noise condition were newly generated for each presentation and faded in and out in the same manner during the experimental procedure. The responses obtained in the random noise condition were not considered in the analysis as this condition was only included to prevent listeners from noise learning. Note that, for a given speech token, the frozen-noise tokens in this experiment were the same across all SNR conditions ("frozen"), in contrast to experiment 1 where different noise tokens were used across SNR conditions ("half-frozen"). SNR conditions of 12, 6, 0, −6, −12, and −15 dB were created by fixing the noise level to 60 dB SPL and adjusting the level of the speech tokens (cf. Sec. II A 2). Each speech token was mixed with the two respective frozen-noise tokens such that the speech token onset was temporally positioned 400 ms after the noise onset. For the random-noise condition, the same was done during the experiment using randomly generated noise waveforms. The clean speech at the respective levels, the frozen-noise tokens, and the mixture of speech and frozen noise tokens were individually stored in ".wav" format at a sampling rate of 44.1 kHz with a resolution 16 bits per sample.

3. Experimental design

As in experiment 1, two control conditions were defined ("Q60" and "Q45"), in which the speech was presented in quiet at 60 and 45 dB SPL, respectively (see also Sec. II A 3). The experiment was split into eight consecutive blocks corresponding to the eight experimental conditions (in the following order: Q60, Q45, SNR = 12, 6, 0, −6, −12, and −15 dB). Each block consisted of a training run followed by the experiment run. For the quiet conditions Q60 and Q45, the training run comprised one presentation of each of the 15 speech tokens; in the experiment run, each of

the speech tokens was presented 5 times, amounting to 75 presentations. The order of presentation was randomized. For the SNR conditions, the training run consisted of 3 presentations of each of the 15 speech tokens, i.e., 45 presentations. The masking noise was newly generated for each presentation during the training run. In the experiment run, each speech token was presented 5 times in each masking-noise condition, i.e., 5 times in frozen noise A, 5 times in frozen noise B, and 5 times in random noise, resulting in a total of 225 presentations. The order of presentation was randomized. One entire experimental block comprised 90 (quiet conditions Q60 and Q45) or 270 (main conditions, SNR: −15. −12, −6, 0, 6, 12 dB), the whole experiment 1800 stimulus presentations. The full experiment lasted approximately 3 h including instruction, training, and breaks.

4. Experimental procedure and apparatus

The listening situation, instructions, interface, and further technical details were the same as in experiment 1, described in Sec. II A 4. The data were converted to proportions of responses in the same manner as described in Sec. II A 4. Five observations per stimulus and individual listener had been obtained; the number of pooled observations was thus 40 (5 observations × 8 listeners).

C. Perceptual distance calculation

To quantify the size of the different source-induced and receiver-related effects, a measure of perceptual distance was applied. Following an approach suggested by Scheidiger and Allen (2013), each response alternative (i.e., each consonant) was considered to represent one dimension in an R-dimensional space (with R denoting the number of response alternatives). In this space, each response pattern was considered as a vector. The perceptual distance between two such response patterns was calculated as the normalized angular distance between two R-dimensional response vectors $\mathbf{x}$ and $\mathbf{y}$,

$$\mathrm{D}[\mathbf{x},\mathbf{y}] = \cos^{-1}\left(\frac{\langle \mathbf{x},\mathbf{y}\rangle}{\|\mathbf{x}\|\cdot\|\mathbf{y}\|}\right)\cdot\frac{100\%}{\frac{\pi}{2}}, \quad (1)$$

where $\langle \mathbf{x},\mathbf{y}\rangle$ denotes the scalar product and $\|\mathbf{x}\|$ and $\|\mathbf{y}\|$ represents the Euclidean norm of the response vectors $\mathbf{x}$ and $\mathbf{y}$, respectively. The response vectors contain the proportions of responses for all R response alternatives (with $\mathrm{R}=15$); thus, the values of the individual coordinates range from 0 to 1 and the angular distance between the two vectors therefore ranges from 0 to $\pi/2$. Normalization by $\pi/2$ and multiplication by 100% yields the normalized angular distance in percent.

The perceptual distance measure was used to describe the amount of perceptual variability induced by the different sources of variability considered in this study. It was calculated between all pairwise combinations of individual listeners' responses that are representative of each factor. For instance, the perceptual influence of across-talker variability can be described using the perceptual distances between all pairs of response vectors obtained with pairs of speech tokens of the same phonetic identity that were spoken by different talkers.

The calculations were performed for each SNR condition separately based on the response vectors obtained with the individual listeners. Depending on the factor, the number of considered response pairs and thus the number of individual distance values varied. For each factor and each SNR condition, a distribution of perceptual distance values (across the considered response pairs) was obtained. As a reference for maximal distance, the perceptual distance *across CVs* ($\mathrm{D_{acrCV}}$) was calculated from the data obtained in experiment 1. To quantify the source-induced variability, the perceptual distances across talkers, within talkers ($\mathrm{D_{acrTalk}}$ and $\mathrm{D_{wtnTalk}}$, both based on experiment 1), and across noise tokens ($\mathrm{D_{acrNoise}}$, based on experiment 2: frozen noise A vs frozen noise B condition) were calculated using response vectors obtained with physically different stimuli of the same phonetic identity. To assess the receiver-related variability, the perceptual distances of the responses across listeners ($\mathrm{D_{acrList}}$, based on experiments 1 and 2) and within listeners ($\mathrm{D_{wtnList}}$, based on experiment 2 test vs retest) were calculated by comparing response vectors obtained with physically identical stimuli. The perceptual distance within listeners represents the listener uncertainty and was thus considered as a baseline for minimal perceptual distance. A detailed description of the perceptual distance calculation is provided in Appendix B.

A more common descriptor of response variability used in related studies (e.g., Miller and Nicely, 1955; Phatak *et al.*, 2008) is the entropy of responses. For comparison with the results obtained based on the perceptual distance measure, the data were also analyzed using the normalized entropy (see Appendix C for details). The perceptual distance may provide an intuitive approach for investigating the perceptual effects of the different sources of variability, whereas the application of the normalized entropy for this purpose is formally less straightforward (see Appendixes B and C).

III. RESULTS

A. Consonant recognition in quiet

The average recognition rate across all 90 speech tokens used in experiment 1 was found to be 99.2% with a standard deviation of 2.7% across CVs for Q60 (at 60 dB SPL presentation level), while the average recognition rate was 96.1% with a standard deviation of 8.5% across CVs for Q45 (at 45 dB SPL). Regarding experiment 2, the analysis showed that the average recognition rate across all 15 speech tokens was 98.8% with a standard deviation of 3.4% across CVs for Q60, while the average recognition rate was 98.2% with a standard deviation of 4.8% across CVs for Q45. All speech tokens used in the two experiments were thus considered sufficiently recognizable in quiet and taken into account for the further analyses.

B. Source-induced variability

To illustrate the source-induced variability in consonant perception, i.e., perceptual differences that occur for physically different stimuli of the same phonetic identity, selected

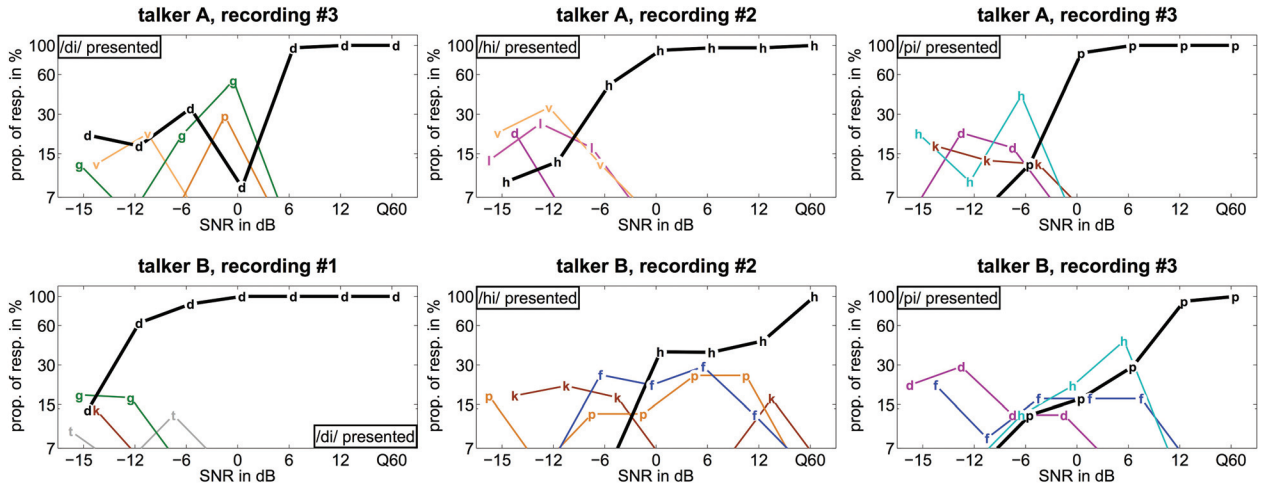

FIG. 1. Across-talker comparison of average confusion patterns for /di/ (left panels), /hi/ (middle panels), and /pi/ (right panels). The upper and lower panels show the results for talker A (male) and talker B (female), respectively. The correct responses are indicated as thick black lines and confusions are shown as thinner lines in different colors; the data points are labeled with the corresponding consonants. Only the four predominant responses are depicted for clarity. A slight horizontal jitter was introduced to the data for better readability. The ordinate is scaled logarithmically to emphasize the confusions. The 7% minimum of the ordinate represents chance level.

example confusion patterns (CPs) are shown for the "average listener," representing the average proportions of responses obtained with eight listeners. The examples illustrate the large observed effect of the considered source variations on consonant recognition and confusions. An analysis of the complete data set follows in Sec. IV.

1. Speech-induced variability

Figure 1 shows average CPs obtained in experiment 1 for the CVs /di/ (left panels), /hi/ (middle panels), and /pi/ (right panels), spoken by the male talker A (top panels) and the female talker B (bottom panels), respectively. The figure illustrates the perceptual effect of *across-talker variability*. For a given speech token, the CPs show the proportions of the four predominant responses as a function of SNR. The proportions of correct responses, denoted as recognition curves, are depicted as thick black lines. The thinner colored lines indicate confusions. The Q60 quiet condition is included as a reference (the rightmost value on the abscissae).

It can be seen that /di/ spoken by talker A (top left panel) is far more confusable (mainly with /gi/) and hence far less recognizable than /di/ spoken by talker B (bottom left panel), particularly at SNRs between −12 and 0 dB. In contrast, an utterance of /hi/ spoken by talker A (top middle panel) was perfectly recognized by the listeners at SNRs down to 0 dB while the same CV spoken by talker B (bottom middle panel) yielded pronounced confusions (with /pi/, /ki/, and /fi/) and thus recognition rates of less than 50% at the same SNRs (0, 6, and 12 dB). Comparably large differences were observed between /pi/ spoken by talker A (top right panel) and /pi/ spoken by talker B (bottom right panel), especially at SNRs of 0 and 6 dB.

Figure 2 shows average CPs obtained in experiment 1 for two different recordings of the CVs /gi/ spoken by the

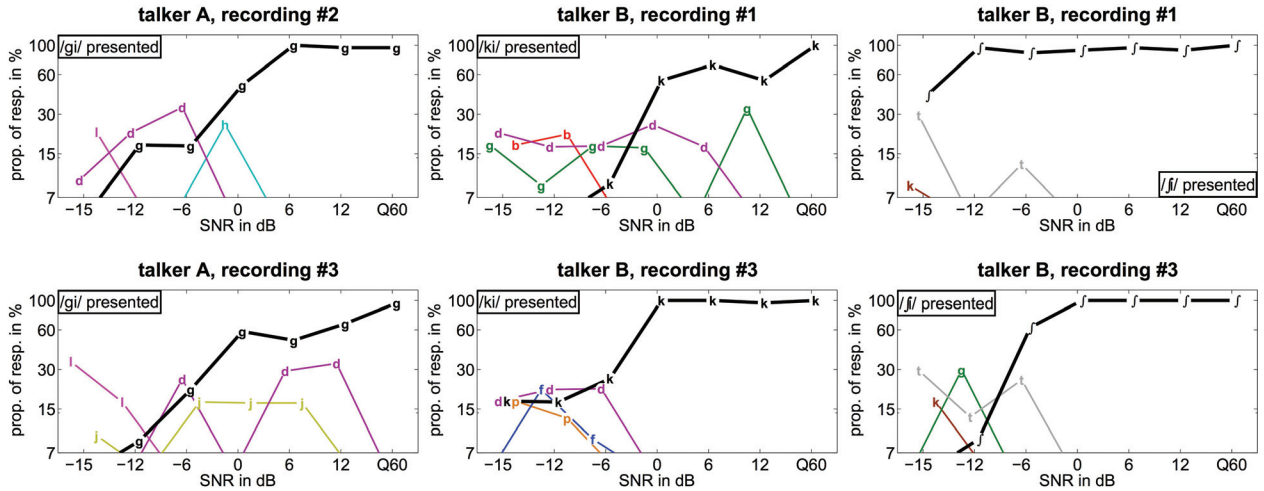

FIG. 2. Within-talker comparison of average confusion patterns for /gi/ (left panels), /ki/ (middle panels), and /ʃi/ (right panels). The upper and lower panels show the results for two different recordings of the same CV, spoken by the same talker (male talker A in the case of /gi/ and female talker B in the cases of /ki/ and /ʃi/). The confusion patterns were obtained as described in Fig. 1.

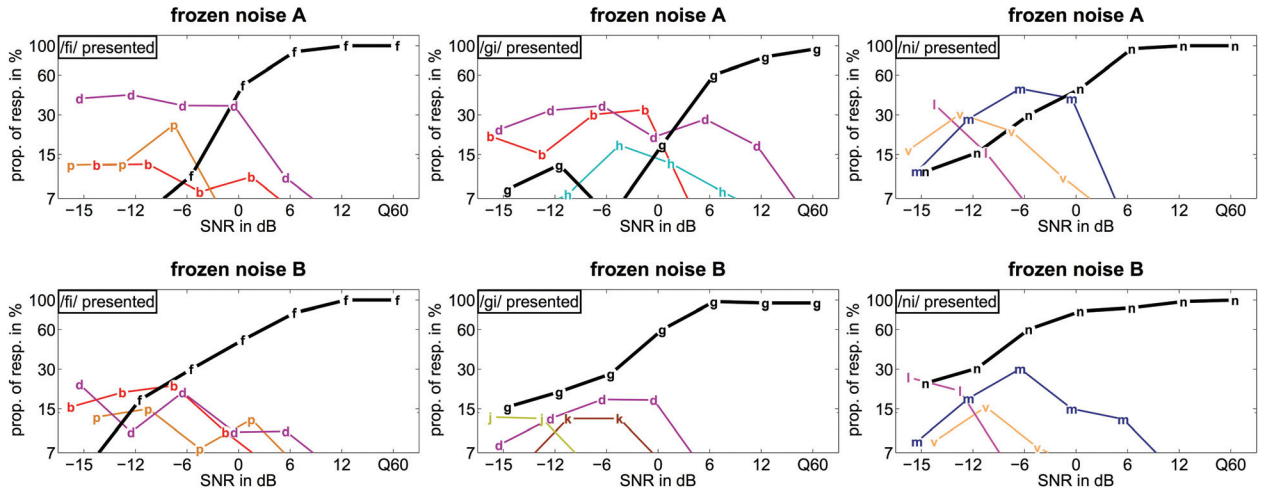

FIG. 3. Across-noise token comparison of average confusion patterns for /fi/ (left panels), /gi/ (middle panels), and /ni/ (right panels). The upper and lower panels show the confusion patterns for the same speech token mixed with different waveforms of frozen noise (top: frozen noise A, bottom: frozen noise B). All speech tokens were spoken by male talker A. The only difference between frozen noise A and frozen noise B was a 100-ms temporal shift. The confusion patterns were obtained as described in Fig. 1.

same male talker A (left panels), /ki/ spoken by the same female talker B (middle panels), and /ʃi/ spoken again by female talker B (right panels). The figure thus illustrates the perceptual effect of *within-talker variability*.

The two different recordings of /gi/ spoken by talker A (left panels) caused large differences in the recognition curves and the confusions, particularly at SNRs of 6 and 12 dB. Similarly, the two recordings of /ki/ spoken by talker B (middle panels) yielded substantially different recognition rates and confusions, particularly at SNRs of 0, 6, and 12 dB. Regarding the two recordings of /ʃi/ spoken by talker B (right panels), it can be seen that this CV was generally detected quite robustly. However, large differences in the recognition rates obtained with the two different recordings were observed for SNRs of −6 and −12 dB.

2. Noise-induced variability

Figure 3 shows average CPs obtained in experiment 2 for the speech tokens /fi/ (left panels), /gi/ (middle panels), and /ni/ (right panels), each presented in frozen noise A (top) and frozen noise B (bottom), respectively. All speech tokens were spoken by the male talker A. Thus, the only difference in the acoustic waveforms of the considered stimulus pairs was a 100-ms temporal shift in the masking-noise waveform.

For the same recording of /fi/ (left panels), the two different noise waveforms led to different CPs. Noise A (top) caused a steeply sloping recognition curve due to a major confusion with /di/ while noise B (bottom) produced a shallower recognition curve as the /di/ confusion was less pronounced. In the case of /gi/ (middle panels), noise A (top) and noise B (bottom) led to substantial differences in the recognition rate at most SNRs as noise A produced different and more pronounced confusions and thus lower recognition rates as compared to noise B. Regarding the results for /ni/ (right panels), noise A (top) yielded a more steeply sloping recognition curve than noise B (bottom). The confusions obtained with the two noise waveforms were the same but much more pronounced for noise A than for noise B.

C. Receiver-related variability

Here, examples of receiver-related variability are shown in terms of selected confusion matrices (CMs). The results demonstrate the observed effect of perceptual differences that occur across listeners and within listeners when no source-induced variability is present, i.e., for physically identical stimuli. An overall analysis of the results follows (Sec. IV).

1. Across-listener variability

Figure 4 shows the across-SNR average of CMs obtained in experiment 2 for four individual listeners. Only the responses obtained in noise B were considered here; thus, the speech and noise waveforms of the stimuli were identical across repeated stimulus presentations and across listeners. The left and right panels show two examples, each comparing the data obtained with two listeners. Each row in the CM reflects the across-SNR average of the proportions of responses obtained for a given speech token mixed with a given noise waveform. The circles indicate the proportions of responses; the filled gray circles show the data obtained with listeners 1 (left) and 3 (right) and the open red circles represent the data obtained with listeners 2 (left) and 4 (right). Thus, the amount of overlap between the filled gray circles and the open red circles indicates the agreement between the responses of listeners 1 and 2 (left) and listeners 3 and 4 (right). The figure hence illustrates the effect of across-listener variability.

Comparing the results of listeners 1 and 2 (left panel), considerable differences can be seen. For example, for /di/ and /ni/, listener 1 showed a larger recognition rate than listener 2, reflected along the diagonal where the filled gray circles exceed the open red circles in size. In contrast, for /fi/ and /mi/, listener 1 showed a smaller recognition rate than listener 2. Regarding confusions, represented by the off-diagonal circles in the CMs, a large variability was found. Some of the major confusions occurred in both listeners (e.g., /di/ confused with

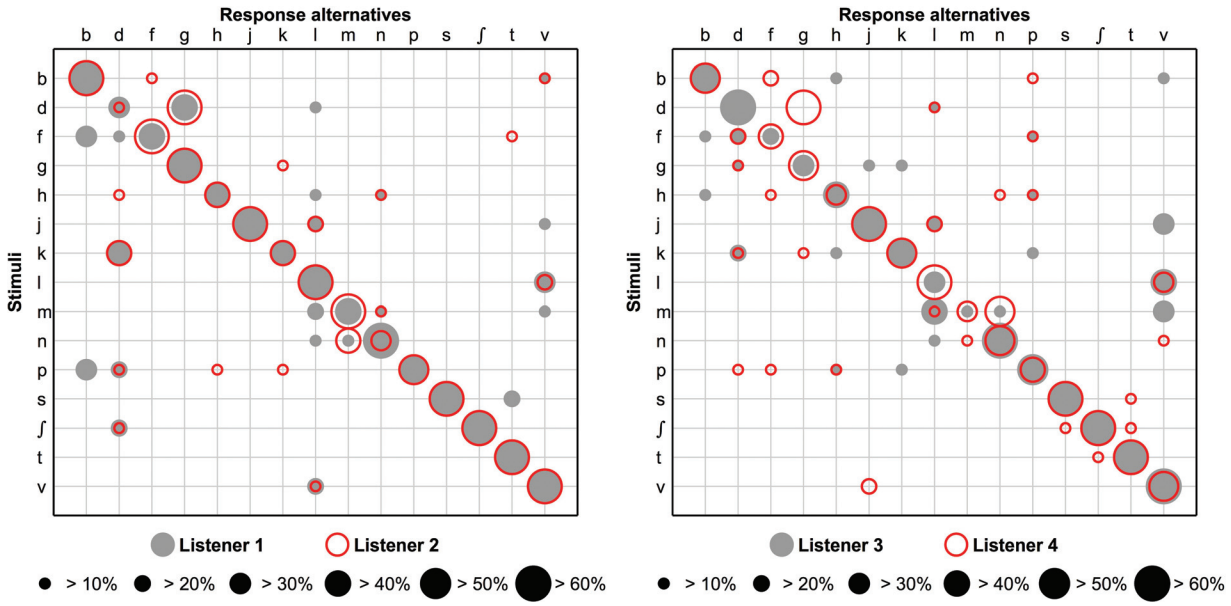

FIG. 4. Across-listener comparison of confusion matrices for the 15 speech tokens used in experiment 2, mixed with frozen noise B. The speech and noise waveforms presented to the individual listeners were identical. For visual clarity, the responses were averaged across SNR conditions. For each stimulus (in each row of the matrix), the size of the circles indicates the proportions of responses for the individual response alternatives (columns of the matrix). Left: responses of listener 1 (gray filled circles) vs listener 2 (red open circles). Right: responses of listener 3 (gray filled circles) vs listener 4 (red open circles).

/gi/). However, the proportions of the individual confusions mostly differed, as indicated by the differences in the size of the overlapping off-diagonal filled gray and open red circles. Furthermore, many distinct confusions made by listener 1 (e.g., /fi/ confused with /bi/) were not made by listener 2 and vice versa.

Comparing the results of listeners 3 and 4 (right panel), the inter-individual differences in the results become even more apparent. The recognition rates (diagonal entries in the CMs) differed for /di/, /fi/, /gi/, /hi/, /li/, /mi/, /ni/, /pi/, and /vi/ (i.e., for nine out of fifteen CVs). Particularly in the case of /di/, listener 3 showed a high recognition rate, while listener 4 selected /gi/ instead of /di/ in about the same number of presentations. Some of the confusions were observed in both listeners, indicated by the overlapping off-diagonal filled gray and open red circles (e.g., /li/ confused with /vi/). However, the proportions of the shared confusions differed and most of the confusions made by listener 3 were not made by listener 4, as indicated by the non-overlapping off-diagonal filled gray and open red circles.

2. Within-listener variability

Figure 5 shows the across-SNR average of CMs obtained in test and retest of experiment 2 for two individual listeners. As above, only the responses obtained in frozen noise B were considered here; the speech and noise waveforms of the stimuli were therefore identical across repeated stimulus presentations and across test and retest. The illustration of the CMs is equivalent to the one used above. However, while Fig. 4 compared results across two pairs of listeners (listener 1 vs 2 and listener 3 vs 4), the left and right panels of Fig. 5 show the comparison of results obtained in test and retest for two individual listeners (listener 1 and listener 3). The figure therefore illustrates the effect of within-listener variability.

Listener 1 (left panel) showed fairly similar recognition rates in test and retest, as can be seen from the overlap of the filled gray and the open red circles along the diagonal. However, for /fi/, /gi/, /ki/, /pi/, and /vi/, the recognition rates were found to be slightly larger in the test (filled gray circles) than in the retest (open red circles). Regarding confusions, it can be seen that most of the major confusions were reproducible since most of the large off-diagonal circles overlap. However, the proportions of the confusions differed slightly across test and retest results, indicated by the differences in the sizes of the filled gray and the open red off-diagonal circles.

Listener 3 (right panel) also showed a large similarity of results obtained in test and retest. The recognition rates were found to be virtually identical, indicated by the perfect overlap of the on-diagonal filled gray and open red circles. Two exceptions were the results for /li/ and /mi/, where the recognition rate in the retest (open red circles) exceeded the recognition rate in the test (filled gray circles). As observed for listener 1, most of the major confusions were reproducible since most of the large off-diagonal circles share a large overlap while the proportions of the confusions partly differed between the test and retest results.

IV. ANALYSIS

The entire data set of the present study was analyzed in terms of source-induced and receiver-related effects using perceptual distance distributions as defined in Sec. II C. Figure 6 shows the mean perceptual distances, in percent, derived from the response variability across CVs (black), across talkers (blue), within talkers (green), across noise tokens (red), across listeners (light gray), and within listeners

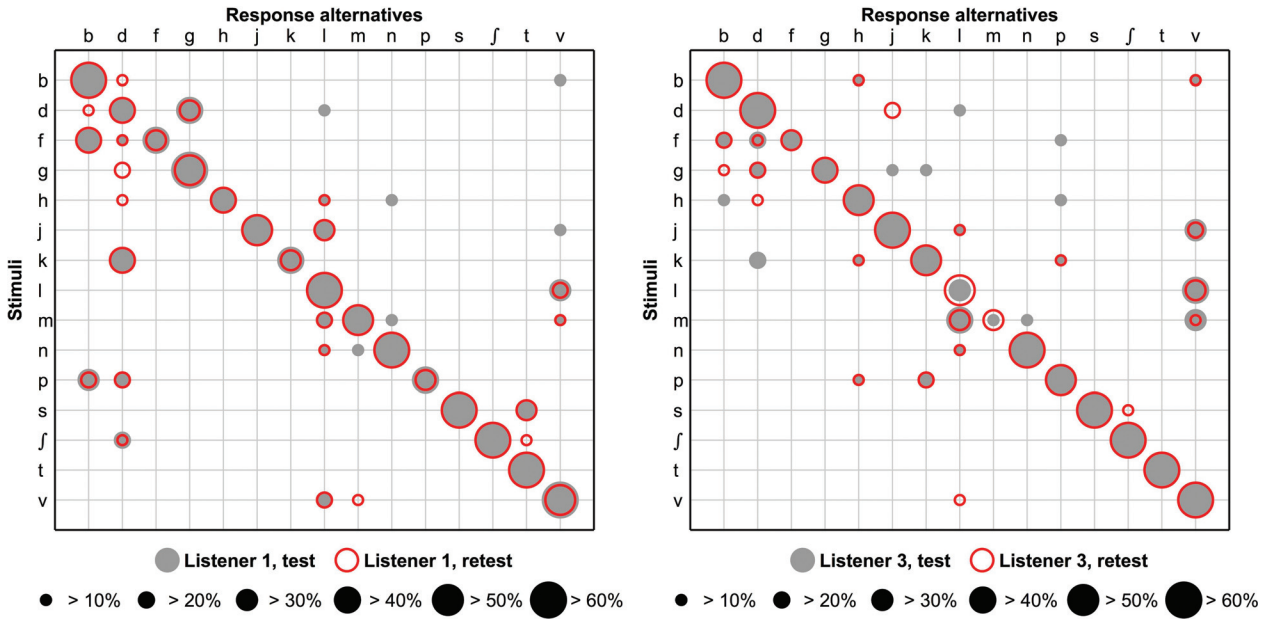

FIG. 5. Within-listener comparison of confusion matrices for the 15 speech tokens used in experiment 2, mixed with frozen noise B. The speech and noise waveforms presented to the listeners in test and retest were identical. For visual clarity, the responses were averaged across SNR conditions. The confusion matrix depiction was obtained as in Fig. 4. Left: responses of listener 1 obtained in test (gray filled circles) and retest (red open circles). Right: responses of listener 3 obtained in test (gray filled circles) and retest (red open circles).

(dark gray), respectively, as a function of SNR. On the left, the average across SNR is shown. The error bars indicate the standard error across the underlying distributions of perceptual distance values obtained with the individual response pairs. The standard errors are proportional to the number of the respective considered response pairs, which varied greatly across the individual sources of variability (across CVs: 30 240; across talkers: 1080; within talkers: 720; across noise: 120; across listeners: 3360; within listeners: 120).

The averages of the perceptual distances across SNR (leftmost bars) provide a good approximation of the size of the perceptual effects induced by the considered sources of variability. The reference for maximal perceptual distance, the perceptual distance across CVs (black bars), confirmed the expected large effect of consonant identity (91%). Regarding the source-induced perceptual distances across stimuli of the same phonetic identity, the largest perceptual distance of 51% was obtained for the across-talker condition (blue bar), followed by the perceptual distance of 47% obtained for the within-talker condition (green bar). This indicates that articulatory differences in utterances of a given talker had a perceptually comparable effect to articulatory differences in utterances of different talkers of different gender. The perhaps most striking observation was that even a slight temporal shift in the waveform of the noise masker mixed with the same speech token produced a considerable effect and led to a perceptual distance of 39% (red bar). Regarding the receiver-related effects, a substantial across-listener effect was found, corresponding to a perceptual distance of 46% for physically identical stimuli (light gray bar). This indicates a large variability in the consonant perception across NH listeners with similar language background. The

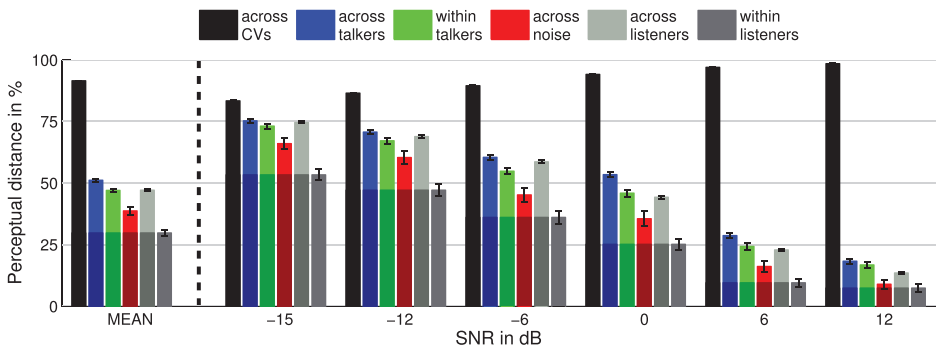

FIG. 6. Mean perceptual distances as a function of SNR and averaged across SNR (left cluster). The error bars represent the standard error across the considered response pairs. As a reference for the maximum occurring perceptual distance, the perceptual distance across different CVs is shown (black bars). Comparing responses to physically different stimuli that share the same phonetic identity, the perceptual distances across talkers (blue bars), within talkers (green bars), and across frozen masking-noise tokens mixed with the same speech token (red bars) are depicted. Comparing responses across physically identical stimuli, the perceptual distances across listeners (light gray bars) and within listeners (dark gray bars) are shown. The shaded areas represent values below the within-listener distance, i.e., below the internal-noise baseline.

across-listener effect was found to be as large as that resulting from within-talker variability (47%, green bar). In other words, the perceptual variability *across listeners* presented with physically identical stimuli was in the range of the perceptual variability in *individual listeners* induced by different speech tokens of the same phonetic type. In contrast, the relatively low perceptual distance within listeners of 30% (dark gray bar) indicated that the individual listeners were able to reproduce their responses fairly reliably.

Two-tailed paired-sample t-tests were performed to verify the statistical significance of the differences observed across the considered conditions. The across-CV reference condition was not considered here. The tests were conducted based on the across-SNR average of the obtained distance distributions. As the sample sizes for the individual conditions differed, with 120 being the minimum sample size, 120 observations were randomly chosen from each sample. The procedure was iterated 10 000 times for convergence and the resulting p-values and t-values were then averaged. A significance level of $\alpha = 0.05$ was assumed and divided by 10 to correct for the ten considered comparisons between the five remaining conditions ($\alpha_{corr} = 0.005$). The results are given in Table I and indicate that all considered conditions were significantly different from each other ($p < 0.005$) except the across-talker, within-talker and across-listener conditions.

The SNR-specific results show that the within-listener perceptual distance (dark gray bars in Fig. 6) increased with decreasing SNR. The lower the SNR, the more challenging was the task and the less reproducible were the responses of the individual listeners obtained with identical stimuli in test and retest. The within-listener response variability represents an intrinsic limitation and was therefore considered as the baseline ("internal noise"). The perceptual distances obtained across talkers (blue bars), within talkers (green bars), across noise tokens (red bars), and across talkers (light gray bars) increased along with—but were well above—the within-listener distance (as indicated by the shaded regions in Fig. 6). The perceptual distance across CVs, reflecting the maximal perceptual distance, showed the opposite trend since responses obtained with stimuli of different phonetic identity were compared: when the task was easy, the perceptual distance across responses obtained with stimuli of different phonetic identity was at ceiling (e.g., /bi/ and /di/ correctly recognized at large SNRs, response vectors thus orthogonal); with decreasing SNR the perceptual distance across these responses decreased as the recognition dropped and the number of confusions increased. Thus, while the perceptual distance across CVs (black bars) represented the largest contribution at all SNRs, it almost reached the level of the CV-specific perceptual distances for an SNR of −15 dB. Disregarding the influence of the listener uncertainty (within-listener distance, dark gray bars), the relation between the CV-specific perceptual distances remained almost the same at all SNRs. Thus, the across-SNR average distances described earlier capture the main effects observed at all SNRs.

TABLE I. T-test results obtained for across-SNR average perceptual distance distributions. Bold numbers indicate p-values < 0.005, with 0.005 being the significance level after Bonferroni correction.

Conditions	t(119)	p
$D_{wtnList}$ vs $D_{acrTalk}$	9.9448	**0.0000**
$D_{wtnList}$ vs $D_{wtnTalk}$	8.2776	**0.0000**
$D_{wtnList}$ vs $D_{acrNoise}$	4.6621	**0.0000**
$D_{wtnList}$ vs $D_{acrList}$	8.2242	**0.0000**
$D_{acrNoise}$ vs $D_{acrTalk}$	5.3434	**0.0000**
$D_{acrNoise}$ vs $D_{wtnTalk}$	3.6554	**0.0031**
$D_{acrNoise}$ vs $D_{acrList}$	3.6789	**0.0040**
$D_{acrTalk}$ vs $D_{wtnTalk}$	−1.6744	0.2038
$D_{acrTalk}$ vs $D_{acrList}$	−1.6130	0.2210
$D_{wtnTalk}$ vs $D_{acrList}$	0.0441	0.5194

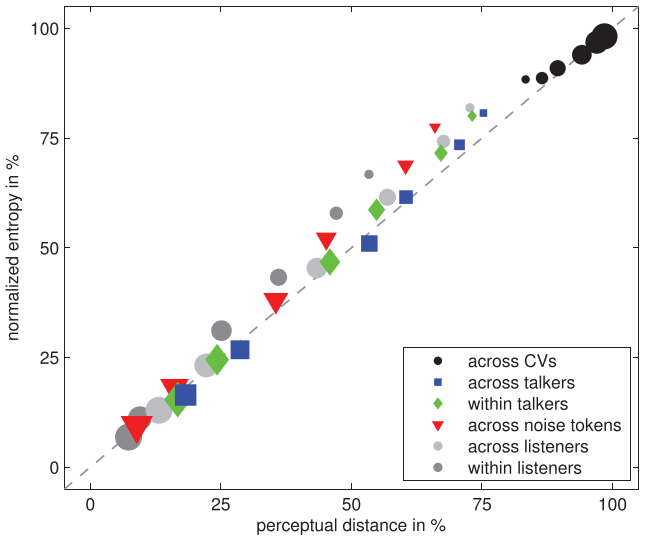

FIG. 7. Scatter plot of perceptual distance in percent (abscissa) versus normalized entropy in percent (ordinate) for the different considered conditions: across CVs (black circles), across talkers (blue squares), within talkers (green diamonds), across noise tokens (red triangles), across listeners (light gray circles), and within listeners (dark gray circles). The perceptual distance and normalized entropy values obtained for the six different SNR conditions (12, 6, 0, −6, −12, and −15 dB) are plotted against each other. The sizes of the respective symbols are proportional to the SNR values. The gray diagonal dashed line represents perfect correlation of perceptual distance and normalized entropy.

Figure 7 shows the relation between the perceptual distance (abscissa) and the normalized entropy (ordinate) by means of a scatter plot. The respective conditions are indicated by different colors and symbols, whereas the different SNRs are indicated by the size of the symbols. For large SNRs, the normalized entropy and the perceptual distance were almost fully correlated, i.e., the large symbols lie on top of the diagonal. For lower SNRs, the normalized entropy slightly exceeded the perceptual distance and the correlation between the two measures thus slightly decreased. Still, the overall correlation was 0.99. The SNR-specific correlation coefficients decreased with decreasing SNR but were all above 0.98. The results obtained based on the perceptual distance are thus supported by the concept of entropy.

V. DISCUSSION

A. Summary of main findings

The present study investigated the effects of different sources of variability in NH listeners' perception of

consonants presented in steady-state masking noise. Two main categories of perceptual variability were defined: source-induced and receiver-related variability. The former describes perceptual differences caused by acoustical differences in stimuli of the same phonetic identity and was subdivided into speech-induced variability (across talkers and within talkers) and noise-induced variability. A special case of source-induced variability is the variability across consonants, which has been considered here as a reference for maximal variability. The latter comprises perceptual differences across listeners and within listeners. To quantify the relative influence of the individual sources of variability, the responses obtained in two experiments were analyzed in terms of example comparisons using a subset of the data and by means of a perceptual distance measure and the entropy of responses using the entire data set.

Regarding the source-induced variability for stimuli of the same phonetic identity, it was shown that the largest perceptual variability was induced by across-talker articulatory differences, closely followed by the effect of within-talker articulatory differences. Furthermore, even a slight temporal shift in the waveform of the steady-state masking noise was found to produce a smaller, yet clearly measurable and statistically significant perceptual effect. Regarding receiver-related variability, the analysis showed that, for physically identical stimuli, the perceptual differences across the NH listeners were very large (in the range of the speech-induced differences). In contrast, the within-listener variability (listener uncertainty) was found to be much smaller, indicating that the reproducibility of the responses for individual listeners was much larger than the agreement between the responses of different listeners. The within-listener variability depended inversely on the SNR, i.e., the "internal noise" (listener uncertainty) was proportional to the "external noise" (acoustic noise).

B. Relation to other studies

In the present study, a large perceptual effect of across-talker articulatory differences was found for identical CVs. This is consistent with other recent studies on consonant perception (e.g., Phatak *et al.*, 2008; Singh and Allen, 2012; Toscano and Allen, 2014), which demonstrated that different speech tokens of the same phonetic identity spoken by different talkers elicit largely different percepts. In contrast, early studies on consonant perception (e.g., Miller and Nicely, 1955; Wang and Bilger, 1973) pooled the responses obtained with different speech tokens of the same phonetic identity spoken by different talkers, thus neglecting the talker-specific perceptual details.

The effect of within-talker articulatory differences was in the present study found to be almost as large as the effect of across-talker articulatory differences. This has otherwise not been reported yet since related studies on consonant perception typically used only one speech token from a given talker for each CV. A within-talker effect was expected given the natural within-talker articulatory variability; however, the authors of the present study did not expect such a prominent effect.

A significant perceptual effect of a temporal shift in the masking noise waveform was found, demonstrating that different white-noise waveforms, mixed with the same speech token, can elicit different speech percepts. Thus, the common assumption in various previous studies (e.g., Miller and Nicely, 1955; Phatak and Allen, 2007; Phatak *et al.*, 2008) of an invariance of consonant perception across steady-state noise realizations cannot be supported by the present study. In fact, the results obtained here suggest that the interaction between a given speech token and the spectro-temporal details of the "steady-state" masking noise waveform matter in the context of microscopic consonant perception. When analyzing responses obtained with individual speech tokens (as in Li *et al.*, 2010; Li *et al.*, 2012; Singh and Allen, 2012; Toscano and Allen, 2014), averaging responses across noise realizations thus appears problematic.

Furthermore, the results of the present study showed that, even for physically identical stimuli, the across-listener perceptual variability is large. The within-listener perceptual variability was found to be clearly smaller. Studies on consonant perception in NH listeners (e.g., Miller and Nicely, 1955; Phatak and Allen, 2007; Phatak *et al.*, 2008; Toscano and Allen, 2014) relied solely on across-listener average data without assessing deviations from the across-listener average due to inter-individual perceptual differences. Toscano and Allen (2014) stated that listeners were highly consistent without providing explicit evidence for this claim. Their analysis was based on consonant recognition only while the analysis performed in the present study also took consonant confusions into account, which may yield different results. Nevertheless, the assumption that consonant perception of NH listeners with similar language background is as consistent across listeners as within listeners is in contrast to the results of the present study. Therefore, across-listener average data should be treated as a population response that is not representative of individual listeners (and vice versa).

C. Implications for the design of consonant perception experiments

The present study demonstrated that all considered differences in the speech token and/or in the noise token led to different consonant percepts. Further, the perceptual variability across NH listeners with the same language background was found to be large. The implications of these findings for the design of consonant experiments largely depend on the goal of the respective study.

If the goal is to "globally" assess consonant perception as a function of consonant identity and SNR, it should be ensured that the described sources of variability (source-induced and receiver-related) do not bias the resulting data. Thus, (i) many speech tokens spoken by different talkers should be considered for each consonant to cover the speech-induced variability, (ii) randomly generated masking noise should be employed to cover the noise-induced variability, and (iii) many listeners should be tested to cover the across-listener perceptual variability. The responses may then be averaged across different speech tokens of the same phonetic identity, different noise waveforms, and different

listeners, yielding an overall pattern of consonant perception as a function of consonant identity and SNR. A more realistic description of the data obtained in such an experiment may be achieved by interpreting the responses obtained with each considered CV as multi-dimensional probability distributions across speech tokens, noise tokens, listeners, and SNR.

In contrast, if the purpose of the study is to investigate which acoustic cues determine a specific confusion pattern, (i) the responses need to be evaluated for each speech token separately (since different speech tokens can elicit different speech percepts), (ii) the combination of speech token and masking-noise token needs to be unique (since the use of randomly generated masking noise mixed with identical speech tokens can elicit different speech percepts in each trial), and (iii) the responses need to be evaluated in individual listeners (due to the substantial perceptual differences across listeners). This level of detail is also needed when assessing effects of individual hearing impairment and hearing-aid signal processing via consonant perception tests. If the above constraints are not respected, the observed results may well be blurred by speech-induced variability, noise-induced variability, and across-listener perceptual variability. Recent detailed studies of consonant cues (Li *et al.*, 2010; Li *et al.*, 2012) indeed analyzed the data for each speech token individually. However, random realizations of steady-state masking noise were used for each trial in the masking experiments and the analyses were performed based on across-listener average data.

D. Implications for consonant perception modeling

So far, no model has been proposed that is able to predict consonant perception in terms of recognition and confusions. The results of the present study may provide some general constraints for microscopic models of speech perception.

If the goal is to predict the average responses for a given consonant and SNR obtained with many speech tokens, many noise realizations, and many listeners, the model's responses should reflect the same *average* outcome measures obtained with the same set of stimuli. Such a "global" model would not be designed to account for the sources of variability considered in the present study, but may account for the effects of consonant identity and SNR. For such an approach, the observations from the present study motivated that the decision-making process in the model back end should incorporate an internal-noise term that scales with the amount of external noise represented in the stimulus.

If a model of consonant perception is targeted towards more details in the consonant perception results, the model needs to reflect all sources of variability. For instance, the fact that a temporal shift in the noise waveform can lead to substantial perceptual differences indicates that a suitable model front end should be sensitive to signal-to-masker phase relations probably already in the peripheral processing of the stimuli. Furthermore, similar to the considerations regarding the "global" model, the observed relation between SNR and within-listener variability suggests that an internal-noise term which scales with the amount of external noise in the stimulus should be incorporated in the model back end. The observed large across-listener perceptual variability represents a major challenge for modeling, since this variability can either arise from differences in the sensory processing in the individual listeners, or from differences at higher-level processes, or both. Such differences may occur even in the case of NH listeners (as considered in the present study), since (i) the applied "criterion" for NH was not very strict, (ii) the audiogram may not be a sufficient descriptor for sensory processing, and (iii) higher-level speech processing may differ across listeners independent of their sensory capabilities, e.g., due to different cognitive abilities. How these inter-individual differences across NH listeners can be quantified and eventually integrated into a modeling framework remains a major challenge.

E. Limitations of the approach

In the present study, several parameters that are known to play a perceptual role were *fixed* to solely focus on specific sources of variability, as it is not feasible to test all possible factors at once. Specifically, the vowel (/i/), the type of consonant-vowel combination (CV) and the spectral shape of the noise (white) were fixed. The same holds for the choice of response alternatives, response method, and the instructions given to the test subjects. The influence of these parameters, which also represent sources of variability, was thus neglected.

The claims made in the present study were based on a set of 90 speech tokens, spoken by two talkers, and presented to two different panels of eight listeners (experiment 1 and experiment 2) and a panel of four listeners (retest of experiment 2), respectively. Therefore, the results may be biased by the choice of speech tokens, talkers, and listeners. Furthermore, the relative size of the reported effects may be different when considering speech samples obtained from natural speech utterances as opposed to isolated syllable productions.

The individual sources of variability investigated in the present study represent *categories* and only provide indications about the relative contributions of these categories (e.g., across-talker articulatory differences) to consonant perception. Thus, it remains unanswered here which specific acoustical properties of the stimuli caused the observed perceptual differences. The three-dimensional deep search method introduced by Li *et al.* (2010) might be one way of addressing this question in terms of a spectro-temporal analysis. Another approach might be to identify the importance of consonant cue regions as a function of audio frequency and modulation frequency, as suggested by Christiansen *et al.* (2007). Further investigations are required to provide an understanding of the relationship between acoustic features in the noisy speech waveform, their internal representation in the auditory system, and the contribution of the different features to robust phoneme recognition.

VI. SUMMARY AND CONCLUSIONS

An experimental approach to investigate the influence of various sources of variability in consonant perception was presented. The study focused on the consonant perception of

NH listeners presented with CVs in white noise at different SNRs. The perceptual variability was split into two main categories, source-induced and receiver-related variability. Using example-based comparisons for the different conditions, and quantifying the observations by means of a measure of perceptual distance, the relative importance of the individual sources of variability was described. Regarding the source-induced variability, the largest effect was found for across-talker articulatory differences, followed by within-talker articulatory differences. Furthermore, even the waveform of the masking noise was shown to induce a significant perceptual effect. In terms of receiver-related effects, a large variability of the responses across listeners was found, whereas the within-listener variability was rather small. Furthermore, the within-listener variability (i.e., the "internal noise") was found to be proportional to the amount of masking noise ("external noise") in the stimulus.

The results from the present study complement current knowledge on consonant perception. It is suggested that, in addition to speech-induced variability, also noise-induced variability as well as across-listener perceptual variability should be taken into account, which has implications for the design of consonant perception experiments and models of consonant perception.

ACKNOWLEDGMENTS

We would like to thank Søren Jørgensen for various discussions about the design of the study and the data analysis, and Christoph Scheidiger for discussions regarding the perceptual distance measure. We also thank two anonymous reviewers for their very helpful and supportive comments. J.Z. is a fellow in the initial training network INSPIRE, which has been funded with support from the European Commission under Contract No. FP7-PEOPLE-2011-290000.

APPENDIX A: DESCRIPTION OF THE DATA SET

The data set considered for the analysis comprised various factors. Table II provides an overview of the dimensionality of the data set. For clarity, the Q45 and Q60 quiet conditions were not considered here. In experiment 1, 3 recordings from each of the 2 talkers of each of the 15 CVs were used. All speech tokens were mixed with white noise at 6 different SNRs and presented to 8 listeners. The listeners had to select one of 16 response alternatives (15 consonants and "I don't know"). Each speech token was presented 3 times at each SNR ($N_{trial1} = 3$). In experiment 2, only 1 speech token was used for each of the 15 CVs. All speech tokens were mixed with 2 different white noise waveforms at 6 different SNRs and presented to 8 listeners. The random masking-noise condition from this experiment was neglected here since it was not used in the analysis. Again, the listeners had to select one of the same 16 response alternatives. Each combination of speech token and noise waveform was presented 5 times at each SNR ($N_{trial2} = 5$). The retest of experiment 2 was conducted with only $N_{list,retest} = 4$ of the 8 listeners.

TABLE II. Overview of the entire data set along with the mathematical notation used for the individual factors.

Factors	Variable name	Experiment 1	Experiment 2
CVs	c	$N_{CV} = 15$	$N_{CV} = 15$
Talkers	τ	$N_{talk} = 2$	$-(1)$
Recordings	ρ	$N_{rec} = 3$	$-(1)$
Masking-noise conditions	η	$-(1)$	$N_{noise} = 2$
SNRs	s	$N_{SNR} = 6$	$N_{SNR} = 6$
Listeners	l	$N_{list} = 8$	$N_{list} = 8$
Response alternatives	—	$N_{resp} = N_{CV} + 1$	$N_{resp} = N_{CV} + 1$
Trials	ν	$N_{trial1} = 3$	$N_{trial2} = 5$

The responses obtained in experiment 1 are denoted as a function $\mathbf{R_I}(c, \tau, \rho, s, l, \nu)$. Accordingly, the responses obtained in experiment 2 are denoted $\mathbf{R_{II}}(c, \eta, s, l, \nu)$. Table II describes the function variables. For each feasible combination of variables, the functions $\mathbf{R_I}$ and $\mathbf{R_{II}}$ return the vectors $\mathbf{r_I} = [r_{I,1}, r_{I,2}, \ldots, r_{I,N_{resp}}]$ and $\mathbf{r_{II}} = [r_{II,1}, r_{II,2}, \ldots, r_{II,N_{resp}}]$, respectively, which have a length of N_{resp}, a value of "1" for the element corresponding to the chosen response, and a value of "0" for all other elements. The last element of these vectors corresponds to the "I don't know" response and all other elements correspond to the 15 consonants provided as response alternatives. The proportions of responses were obtained by distributing the "I don't know" responses evenly across the 15 other response alternatives, summing the responses across all trials, and finally dividing by the number of trials. For the data $\mathbf{R_I}$ obtained in experiment 1, the conversion of the responses obtained with a given CV, talker, recording, SNR, and listener is expressed as

$$p_{I,i} = \frac{1}{N_{trial1}} \sum_{\nu=1}^{N_{trial1}} r_{I,i}(\nu) + \frac{r_{I,N_{resp}}(\nu)}{N_{CV}}, \quad i = 1, 2, \ldots, N_{CV}. \tag{A1}$$

For the data $\mathbf{R_{II}}$ obtained in experiment 2, the conversion of the responses obtained with a given CV, masking-noise waveform, SNR, and listener is provided by

$$p_{II,i} = \frac{1}{N_{trial2}} \sum_{\nu=1}^{N_{trial2}} r_{II,i}(\nu) + \frac{r_{II,N_{resp}}(\nu)}{N_{CV}}, \quad i = 1, 2, \ldots, N_{CV}. \tag{A2}$$

The resulting vectors $\mathbf{p_I} = [p_{I,1}, p_{I,2}, \ldots, p_{I,N_{CV}}]$ and $\mathbf{p_{II}} = [p_{II,1}, p_{II,2}, \ldots, p_{II,N_{CV}}]$ contain the respective proportions of responses and are summarized as the functions $\mathbf{P_I}(c, \tau, \rho, s, l)$ and $\mathbf{P_{II}}(c, \eta, s, l)$, representing the proportions of responses obtained in the two experiments.

APPENDIX B: CALCULATION OF PERCEPTUAL DISTANCE

The perceptual distance measure is defined in Sec. II C. The perceptual distance *across CVs* was calculated across all response pairs obtained with speech tokens of different phonetic identity based on $\mathbf{P_I}$,

$$\mathrm{D_{acrCV}}(s,l,\tau,\tau',\rho,\rho',\delta) = \mathrm{D}[\mathbf{P_I}(c,\tau,\rho,s,l),\ \mathbf{P_I}(c',\tau',\rho',s,l)], \quad \text{(B1)}$$

where $c=[1,\mathrm{N_{CV}}-1]$, $c'=[c+1,\mathrm{N_{CV}}]$, $\tau=[1,\mathrm{N_{talk}}]$, $\tau'=[1,\mathrm{N_{talk}}]$, $\rho=[1,\mathrm{N_{rec}}]$, and $\rho'=[1,\mathrm{N_{rec}}]$. The variable $\delta=[1,\mathrm{N}_\delta]$ describes all $\mathrm{N}_\delta=\Sigma_{n=1}^{\mathrm{N_{CV}}-1}n$ possible combinations of CV identities. For each SNR, the across-CV distances between $\mathrm{N_{list}}\cdot\mathrm{N_{talk}^2}\cdot\mathrm{N_{rec}^2}\cdot\mathrm{N}_\delta=30\,240$ response vector pairs were calculated.

The perceptual distance *across talkers* was calculated across all response pairs obtained with speech tokens of the same phonetic identity spoken by different talkers, based on $\mathbf{P_I}$,

$$\mathrm{D_{acrTalk}}(s,l,c,\rho,\rho') = \mathrm{D}[\mathbf{P_I}(c,\tau,\rho,s,l),\ \mathbf{P_I}(c,\tau',\rho',s,l)], \quad \text{(B2)}$$

where $\tau=1$ and $\tau'=2=\mathrm{N_{talk}}$, $\rho=[1,\mathrm{N_{rec}}]$, and $\rho'=[1,\mathrm{N_{rec}}]$. For each SNR, the across-talker distances between $\mathrm{N_{list}}\cdot\mathrm{N_{CV}}\cdot\mathrm{N_{rec}^2}=1080$ response vector pairs were calculated.

The perceptual distance *within talkers* was calculated across all response pairs obtained with different speech tokens of the same phonetic identity, spoken by the same talker, based on $\mathbf{P_I}$,

$$\mathrm{D_{wtnTalk}}(s,l,c,\tau,\delta) = \mathrm{D}[\mathbf{P_I}(c,\tau,\rho,s,l),\ \mathbf{P_I}(c,\tau,\rho',s,l)], \quad \text{(B3)}$$

where $\rho=[1,\mathrm{N_{rec}}-1]$ and $\rho'=[\rho,\mathrm{N_{rec}}]$. The variable $\delta=[1,\mathrm{N}_\delta]$ describes all $\mathrm{N}_\delta=\Sigma_{n=1}^{\mathrm{N_{rec}}-1}n$ possible combinations of recordings from a given talker. For each SNR, the within-talker distances between $\mathrm{N_{list}}\cdot\mathrm{N_{CV}}\cdot\mathrm{N_{talk}}\cdot\mathrm{N}_\delta=720$ response vector pairs were calculated.

The perceptual distance *across noise tokens* was calculated across all response pairs obtained with identical speech tokens mixed with two different frozen noise tokens based on $\mathbf{P_{II}}$,

$$\mathrm{D_{acrNoise}}(s,l,c) = \mathrm{D}[\mathbf{P_{II}}(c,\eta,s,l),\ \mathbf{P_{II}}(c,\eta',s,l)], \quad \text{(B4)}$$

where $\eta=1$ and $\eta'=2=\mathrm{N_{noise}}$. For each SNR, the across noise-token distances between $\mathrm{N_{list}}\cdot\mathrm{N_{CV}}=120$ response vector pairs were calculated.

The perceptual distance *across listeners* was calculated across all response pairs obtained with physically identical stimuli but different listeners based on $\mathbf{P_I}$ and $\mathbf{P_{II}}$,

$$\mathrm{D_{acrList,I}}(s,c,\tau,\rho,\delta) = \mathrm{D}[\mathbf{P_I}(c,\tau,\rho,s,l),\ \mathbf{P_I}(c,\tau,\rho,s,l')], \quad \text{(B5)}$$

$$\mathrm{D_{acrList,II}}(s,c,\eta,\delta) = \mathrm{D}[\mathbf{P_{II}}(c,\eta,s,l),\ \mathbf{P_{II}}(c,\eta,s,l')], \quad \text{(B6)}$$

where $l=[1,\mathrm{N_{list}}-1]$ and $l'=[l,\mathrm{N_{list}}]$. The variable $\delta=[1,\mathrm{N}_\delta]$ describes all $\mathrm{N}_\delta=\Sigma_{n=1}^{\mathrm{N_{list}}-1}n$ possible combinations of listeners. For each SNR, the across-listener distances between $\mathrm{N_{CV}}\cdot\mathrm{N_{talk}}\cdot\mathrm{N_{rec}}\cdot\mathrm{N}_\delta=2520$ response vector pairs were calculated from $\mathbf{P_I}$ and between $\mathrm{N_{CV}}\cdot\mathrm{N_{noise}}\cdot\mathrm{N}_\delta=840$ response vector pairs from $\mathbf{P_{II}}$. $\mathrm{D_{acrList,I}}$ and $\mathrm{D_{acrList,II}}$ were combined to $\mathrm{D_{acrList}}$ (comprising $2520+840=3360$ distance values per SNR).

The average perceptual distance *within listeners* was calculated across all response pairs obtained with physically identical stimuli and identical listeners in test and retest of experiment 2, i.e., based on $\mathbf{P_{II,test}}$ and $\mathbf{P_{II,retest}}$,

$$\mathrm{D_{wtnList}}(s,\eta,c,l) = \mathrm{D}[\mathbf{P_{II,test}}(c,\eta,s,l),\mathbf{P_{II,retest}}(c,\eta,s,l)], \quad \text{(B7)}$$

where $l=[1,\mathrm{N_{list,retest}}]$. For each SNR, the within-listener distances between $\mathrm{N_{list,retest}}\cdot\mathrm{N_{CV}}\cdot\mathrm{N_{noise}}=120$ response vector pairs were calculated.

APPENDIX C: CALCULATION OF NORMALIZED ENTROPY

In analogy to the calculation of the perceptual distance, the data were also analyzed in terms of entropy. The entropy specifies the amount of variability in a given response vector. Here, the normalized entropy was used, which was defined as

$$H_{norm}(\mathbf{p}) = \frac{100\%}{\log_2(\min[\mathrm{R},\mathrm{N}])}\cdot\sum_{i=1}^{\mathrm{R}}\mathrm{p}_i\log_2\left(\frac{1}{\mathrm{p}_i}\right),\quad \forall\,\mathrm{p}_i>0, \quad \text{(C1)}$$

where $\mathbf{p}=[\mathrm{p}_1,\mathrm{p}_2,\ldots,\mathrm{p}_\mathrm{R}]$, p_i is the proportion of response alternative i, R denotes the number of response alternatives, and N represents the number of observations. The denominator is the theoretical entropy maximum $H_{\max}=\log_2(\min[\mathrm{R},\mathrm{N}])$. Division by H_{max} thus normalizes the entropy to a range from 0 to 1; multiplication by 100% yields the normalized entropy in percent.

The normalized entropy describes the perceptual variability *for a given* response vector whereas the perceptual distance measure represents a comparison between *a pair* of response vectors. To quantify the effect of the different sources of variability based on the normalized entropy, the perceptual variability induced by the different sources of variability therefore had to be contained in individual response vectors. To obtain such response vectors, the raw data $\mathbf{R_I}$ and $\mathbf{R_{II}}$ were converted to proportions of responses obtained (on a trial-by-trial basis) with (1) different CVs, (2) identical CV but different talkers, (3) identical CV and talker but different recordings, (4) identical speech token but different masking-noise waveforms, (5) identical stimulus but different listeners, (6) identical stimulus and listener but test and retest results. The "I don't know" responses were here attributed to randomly chosen response alternatives. The entropy is sensitive to differences in N, the number of observations. To avoid an obscured across-condition comparison, it was therefore necessary to have the same basic number of observations for all the considered conditions that were to be compared. As this was not the case in the data set (e.g.,

$N_{CV} = 15$ but $N_{talk} = 3$), a "fair" entropy comparison could only be obtained using random processes and iterating them many times for the resulting entropy to converge to its true value. The effective number of observations was set to $N = 3$, which corresponds to N_{trial1} and N_{rec}. To illustrate the procedure, three examples are given that represent the cases (1) $N_{factor} > N$, (2) $N_{factor} < N$, and (3) $N_{factor} = N$.

(1) $N_{factor} > N$. The normalized entropy *across CVs* ($N_{CV} = 15$), E_{acrCV}, was calculated based on $\mathbf{R_I}$. For each SNR condition and for each listener, three vectors $\mathbf{r_I}$ obtained in three randomly chosen trials with three different randomly chosen CVs were collected, summed, and converted to a proportions of response vector $\mathbf{p}$ via division by $N = 3$. In order for the random processes to converge, the procedure was iterated 1000 times. The normalized entropy was calculated and eventually averaged across listeners and iterations.

(2) $N_{factor} < N$. The normalized entropy *across talkers* ($N_{talk} = 2$), $E_{acrTalk}$, was calculated based on $\mathbf{R_I}$. For each SNR condition, for each CV, and for each listener, three vectors $\mathbf{r_I}$ obtained in three randomly chosen trials with (1) one randomly chosen recording of talker A, (2) one randomly chosen recording of talker B, and (3) one recording of talker A or talker B (randomly chosen from the residual recordings) were collected, summed, and converted to a proportions of response vector $\mathbf{p}$ via division by $N = 3$. The procedure was iterated 1000 times for convergence. The normalized entropy was calculated and averaged across CVs, listeners, and iterations.

(3) $N_{factor} = N$. The normalized entropy *within talkers* ($N_{rec} = 3$), $E_{wtnTalk}$, was calculated based on $\mathbf{R_I}$. For each SNR condition, for each CV, for each talker, and for each listener, three vectors $\mathbf{r_I}$ obtained in three randomly chosen trials with the three different recordings spoken by the respective talker were collected, summed, and converted to a proportions of response vector $\mathbf{p}$ via division by $N = 3$. The procedure was iterated 1000 times for convergence. The normalized entropy was calculated and averaged across CVs, talkers, listeners, and iterations.

Similar calculations were performed to obtain the normalized entropy *across noise tokens* ($E_{acrNoise}$), *across listeners* ($E_{acrlist}$), and *within listeners* ($E_{wtnlist}$).

Allen, J. B. (**1994**). "How do humans process and recognize speech?," IEEE Trans. Speech Audio Process. **2**(4), 567–577.

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

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., and Wellekens, C. (**2007**). "Automatic speech recognition and speech variability: A review," Speech Commun. **49**, 763–786.

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), 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*.

Cutler, A., Weber, A., Smits, R., and Cooper, N. (**2004**). "Patterns of English phoneme confusions by native and non-native listeners," J. Acoust. Soc. Am. **116**(6), 3668–3678.

Fletcher, H., and Galt, R. H. (**1950**). "The perception of speech and its relation to telephony," J. Acoust. Soc. Am. **22**(2), 89–151.

French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**(1), 90–119.

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

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.

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.

Mullenix, J. W., Pisoni, D. B., and Martin, C. S. (**1989**). "Some effects of talker variability on spoken word recognition," J. Acoust. Soc. Am. **85**(1), 365–378.

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.

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.

Phatak, S. A., Yoon, Y.-S., Gooler, D. M., and Allen, J. B. (**2009**). "Consonant recognition loss in hearing impaired listeners," J. Acoust. Soc. Am. **126**(5), 2683–2694.

Scheidiger, C., and Allen, J. B. (**2013**). "Effects of NALR on consonant-vowel perception," in *the 4th International Symposium on Auditory and Audiological Research (ISAAR-2013)*, Nyborg, Denmark.

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.

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.

Trevino, A., and Allen, J. B. (**2013**). "Within-consonant perceptual differences in the hearing impaired ear," J. Acoust. Soc. Am. **134**(1), 607–617.

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.