Consonant and Vowel confusions in speech-weighted noise

Sandeep Phatak, Jont Allen

Beckman Institute for Advance Science and Technology
University of Illinois at Urbana-Champaign

Abstract

Confusion analysis for closed-set recognition of 64 consonant-vowel (CV) sounds, spoken by 18 talkers, is presented. In presence of speech-weighted noise the confusions patterns of the CV syllables categorize into three sets, depending on the consonant. The consonant confusions correlate closely with the high frequency spectra of the consonants while the vowel confusions correlate with the vowel durations and the second formant frequencies.

Index Terms: Speech perception, Confusion analysis.

1. Introduction

Analyzing the consonant and vowel confusions made by listeners provides insight into the auditory processing of speech sounds. When combined with a spectro-temporal analysis of the stimuli, the confusion analysis forms a framework for finding the underlying *perceptual features* or the *events* in speech [1]. Events are defined as the features, extracted by the human auditory system, that forms the basis for perception of different speech sounds. Humans appear to use these events to parameterize, learn and recognize different speech sounds. The ability of the auditory system to extract these features makes the human speech recognition highly robust to noise, as compared to the machine recognizers [2]. Confusions are a result of the perceptual parameterization of speech sounds and therefore confusion analysis is essential for decoding the process of human speech recognition.

Averaging the data across SNR [3] or using very few talkers or listeners [4] reduces the utility of confusion matrix (CM) data from some of the past experiments. Also, the speech and noise signals used in the past experiments are not available today. Without an analysis of signals, the CM data is not sufficient to reverse engineer the human speech recognition. Therefore, in order to build a database of confusion data for a commercially available recorded database, a Miller-Nicely ("MN16-55") [5] type experiment was conducted at the University of Illinois ("UIUCs04"). Since UIUCs04 was inspired by MN16-55, the primary goal was to analyze the consonant confusions. The purpose of choosing multiple vowels was to analyze the extent of the effect of vowels on the consonant confusion patterns.

2. Methods

A subset of isolated CV syllables from the LDC-2005S22 corpus, consisting of the same 16 consonants used by Miller and Nicely [5] and 4 vowels - /ɑ/, /ɛ/, /ɪ/ and /æ/. Each of the 64 CVs (16 C $\times$ 4 V) was spoken by 14 talkers. The syllables were presented in quiet condition as well as in presence of a speech-weighted noise at five different signal-to-noise ratios (SNR). The data collection procedure is described in detail in [6].

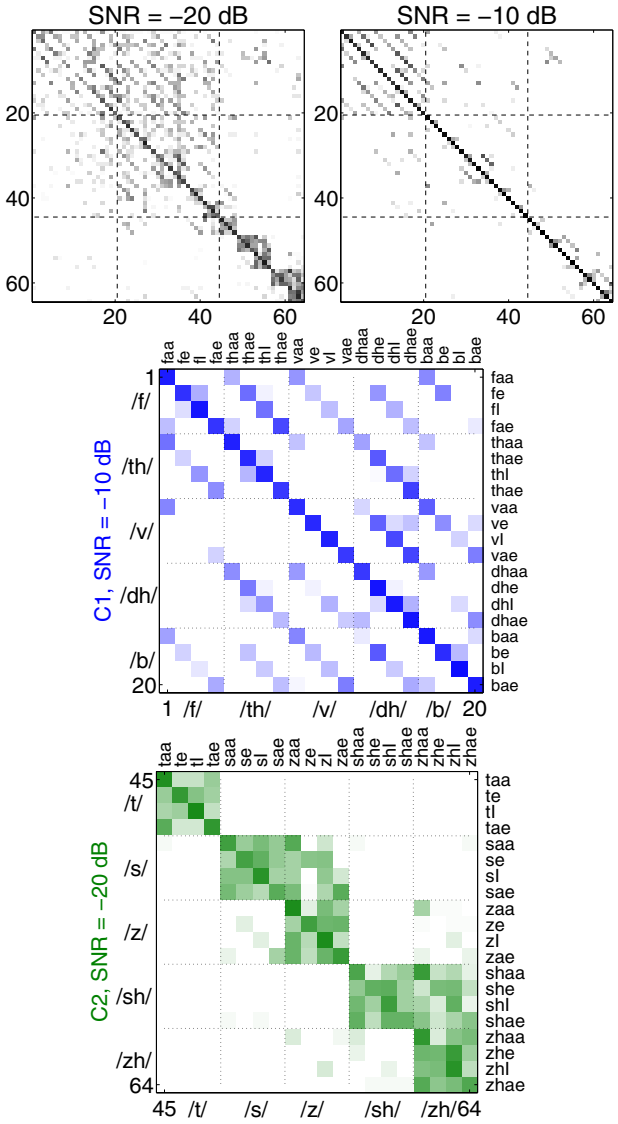

Figure 1: *Top row:* The gray-scale images of the CMs at -20 dB and -10 dB SNR. The gray-scale intensity is proportional to the log of the value of each entry in the row-normalized CM, with black color representing unity and white color representing the chance performance (1/64). Dashed lines separate sets C1 = {/f/, /θ/, /v/, /ð/, /b/}, C3 = {/p/, /g/, /m/, /n/, /k/, /d/} and C2 = {/t/, /s/, /z/, /ʃ/, /ʒ/}, in that order, from left to right and top to bottom. The sequence of the CV sounds is shown in the blow-ups of the top left corner of CM (i.e. set C1) at -10 dB SNR (*Center*) and bottom right corner of CM (i.e. set C2) at -20 dB SNR (*Bottom*).

Fourteen L1=English listeners completed the experiment. Ten "High Performance" (HP) listeners (i.e. listeners with scores greater than 85% in quiet, and greater than 10% correct at -22 dB SNR) formed a homogeneous group. The responses of the remaining four "Low Performance" (LP) listeners were not used for analysis. The responses to the utterances with more than 20% recognition error in quiet condition were also not used for analysis.

3. Results

3.1. Confusion Analysis

Fig. 1 shows the 64×64 syllable CMs at two different SNRs, displayed as gray-scale images. The rows and columns of the CM are arranged such that four CVs having the same consonant are consecutively placed with /ɑ/, /ɛ/, /ɪ/ and /æ/ as the order of vowels, denoted in the figure by the labels /aa/, /e/, /I/ and /ae/, respectively. The labels /dh/, /th/, /sh/ and /zh/ are used for consonants /ð/, /θ/, /ʃ/ and /ʒ/, respectively. With the particular sequence of consonants shown in Fig. 1, two distinct structures can be observed in the CM images - (i) dark lines parallel to the diagonal and (ii) dark blocks on the diagonal. Every fourth line parallel to the diagonal represents a consonant confusion and correct vowel, while a 4 × 4 block around diagonal represents vowel confusions with correct consonant. Based on these two structures, the CV sounds could be grouped into three categories, depending on the consonant in the CV sound (i.e. each category contained all four vowels, but specific consonants). For example, CV sounds with consonants C1 = {/b/, /f/, /θ/, /v/, /ð/} showed more consonant errors (parallel-lines), while those with consonants C2 = {/t/, /s/, /ʃ/, /z/, /ʒ/} showed more vowel errors (block-diagonal). The parallel line structure was observed above -20 dB SNR and all the way up to the quiet condition, but it was smeared below -20 SNR. The CV sounds in set C1, which showed this type of confusion structure, had low overall recognition scores that reached the chance level probability (1.56%) at -22 dB SNR. On the other hand, the block-diagonal structure for set C2 was observed at -22 dB and -20 dB SNR, but was much less obvious above -16 dB SNR. The CV sounds in C2 also showed very high recognition scores that were greater than 10%, even at -22 dB SNR. The remaining CV sounds (set C3) showed a complex combination of both types of structures.

The CVs in the three groups are separated by dashed lines in the CM images, as shown in Fig. 1. Note that many confusions are asymmetric. CVs in C1 are highly confused with those in set C3, and occasionally with those in set C2 at low SNRs. However, CVs in set C3 are confused with those in set C1 but not with those in set C2. The CVs in set C2 are hardly confused with CVs in other two sets. Within set C2, there are asymmetric confusions between /s/-/z/ and /ʃ/-/ʒ/.

3.2. Consonant Confusions

When the syllables were scores only for consonants $[P(\mathcal{C}_h|\mathcal{C}_s\mathcal{V}_s)]$, it was observed that the confusion patterns of consonants /f/, /θ/, /v/, /ð/ (all in C1) and /m/ (set C3) depend on the spoken vowel $\mathcal{V}_s$. Specifically, the consonant confusions for these five consonants were significantly different, with no strong competitor, when the following vowel was /ɑ/. For the other three vowels, the confusion patterns did not vary significantly. Set C1 consonants have scores poorer than the vowel scores and hence the confusion patterns for C1 consonants are more likely to be affected by the following vowel. On the other hand, C2 consonants have scores better than the vowels and the C2 confusion patterns are least likely to be affected by $\mathcal{V}_s$.

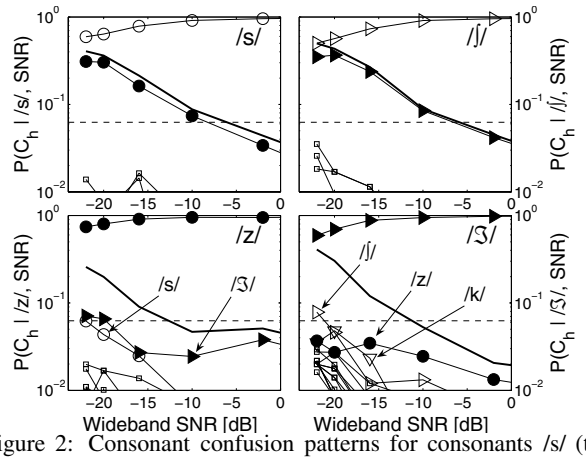

Figure 2: Consonant confusion patterns for consonants /s/ (top left), /ʃ/ (top right), /z/ (bottom left) and /ʒ/ (bottom right).

Since the consonant confusions for four C2 consonant /s/, /ʃ/, /z/ and /ʒ/ are independent of the spoken vowel, their consonant scores are averaged across $\mathcal{V}_s$. Figure 2 shows the corresponding four rows of the vowel-independent consonant CM, $P(\mathcal{C}_h|\mathcal{C}_s)$. The /s/-/z/ and /ʃ/-/ʒ/ confusions are highly asymmetric (Fig. 1, set C2). The total error in recognizing unvoiced consonants /s/ and /ʃ/ can be accounted by the confusions with the voiced consonants /z/ and /ʒ/, respectively, whereas /z/ and /ʒ/ have multiple competitors that contribute to the total error. Thus the asymmetry is biased towards the voiced consonants /z/ and /ʒ/, i.e. these two are the preferred choices in /s/-/z/ and /ʃ/-/ʒ/ confusions in speech-weighted noise. The asymmetric parts of the confusion probability are as high as 0.13 and 0.14 for /s/-/z/ and /ʃ/-/ʒ/ confusions, respectively. This asymmetry is slightly greater than the largest asymmetry found in the consonant CM of MN16-55, which was 0.1 for a different consonant confusion [7].

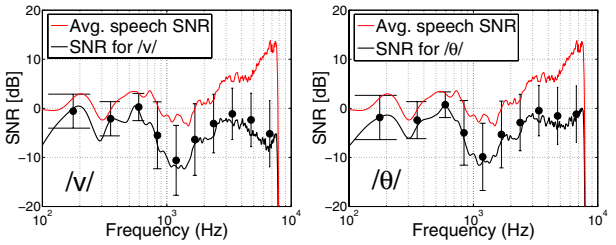

Figure 3: Average SNR spectra (thick solid line) in dB for consonants /v/ and /s/, when the wideband SNR is 0 dB. The thin solid line shows the SNR spectrum for average speech, while the circles and the error-bars show the means and the standard deviations in the values narrow-band SNR estimated in octave bands below 500 Hz and half-octave bands above 500 Hz.

The nature of the confusions among the three consonant sets correlate with the SNR spectrum of the consonants. The SNR spectrum for a consonant is the ratio of power spectral density (PSD) of that consonant to the PSD of speech-weighted noise. To estimate the PSD of a consonant, the PSD of all CV utterances with the given consonant were averaged. Such an average would

practically average out the spectral variations due to different vowels and enhance the consonant spectrum. The thick solid lines in Fig. 3 show the SNR spectra, i.e. SNR as a function of frequency, for two consonants /v/ (set C1) and /s/ (set C2).

The SNR spectrum for consonants in set C1 is lower at high frequencies while those for the C2 consonants increased sharply, reaching above 15 dB for some utterances. This explains the difference between the recognition scores of C1 and C2 consonants. The recognition scores of C2 consonants were above 50% even at -22 dB wideband SNR, while the C1 consonants had relatively greater masking under speech-weighted noise and showed large consonant confusions for SNR $\leq$ -10 dB. This observation is further elaborated in Sec. 3.4.

The SNR spectrum for C3 consonants (not shown) are similar to C1 consonants, having a relatively low SNR at high frequencies. This is consistent with the confusions observed in Fig. 1, which show that the C1 consonants are confused with C3 consonants, but C2 consonants are rarely confused with C1 and C3 consonants.

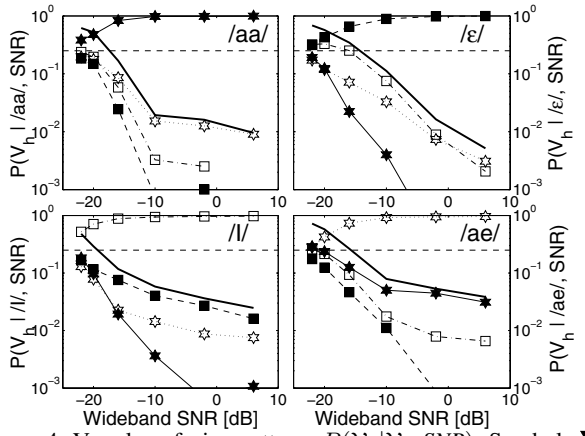

Figure 4: Vowel confusion patterns $P(\mathcal{V}_h|\mathcal{V}_s, SNR)$. Symbols ★, ■ , □ and ✡denote vowels /ɑ/, /ɛ/, /ɪ/ and /æ/, respectively.

3.3. Vowel Confusions

When the syllables are scores only for vowels $[P(\mathcal{V}_h|\mathcal{C}_s\mathcal{V}_s)]$, the vowel confusions were found to be independent of the spoken consonant $\mathcal{C}_s$. Therefore the vowel scores were pooled across $\mathcal{C}_s$ to analyze vowel confusions $[P(\mathcal{V}_h|\mathcal{V}_s)]$ (Fig. 4). At very low SNR values, all the diagonal and the off-diagonal entries in the 4 $\times$ 4 vowel CM converge to the chance level performance of detecting one of the four vowels (25%). The recognition score for each of the four vowels was greater than 30% at -22 dB SNR and was not low enough to see clear groupings with local maxima (SNR_g), with an exception of $P(/\text{ɪ}/|/\text{ɛ}/)$ (top right panel, Fig. 4). However, the off-diagonal entries show some interesting behavior at scores that are an order of magnitude smaller than the chance level. Due to the very large row sums (1700 to 2000 responses) in the vowel CM, the data variability is relatively small resulting in the curves that are distinct and smooth, even at such low values.

At very low SNR, each vowel seemed to be equally confused with the other three vowels, except for /ɛ/, which clearly formed a group with /ɪ/ ($SNR_g \approx$ -20 dB, top right panel of Fig. 4). But as the SNR increased, /ɛ/ became equally confused with /ɪ/ and /æ/, though the total number of confusions decreased. For the other three vowels, the curves of off-diagonal entries separated from each other, showing a clear ordering in the confusability. Above

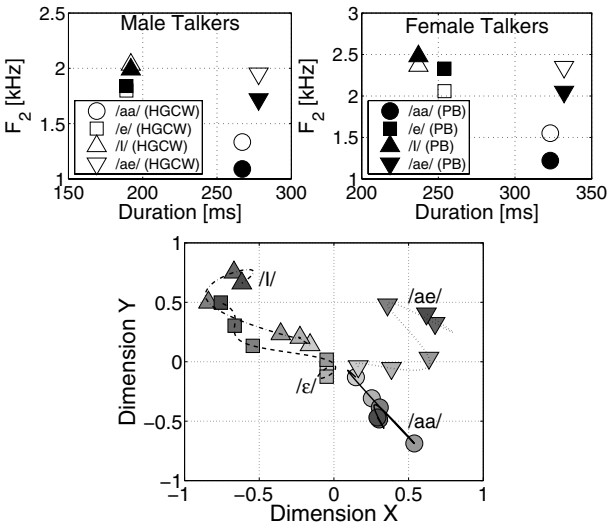

Figure 5: *Top Row:* Plots of the average values of the second formant frequency (F_2) of vowels vs the vowel durations for male (left panel) and female talkers (right panel). The values of the duration are from Hillenbrand *et. al.* [8] (HGCW), while the values of F_2 are from HGCW (hollow symbols) as well as Peterson and Barney [9] (PB) (filled symbols). *Bottom:* 2D projection of the vowel clusters in 3D eigenspace that matches the vowel locations in F_2 vs. Duration space for male talkers. The gray-scale intensity of the symbols show the six SNR levels, with the lightest corresponding to -22 dB SNR and the darkest corresponding to the quiet condition. The lines indicate paths traced by the vowels in the 2-D plane of projection, as the SNR decreases.

-10 dB, /æ/ and /ɑ/ emerged to be the strongest competitors of each other, with /ɪ/ being the next stronger competitor and /ɛ/ being the weakest competitor for both vowels. The vowel /ɛ/ was the strongest competitor of /ɪ/ above -20 dB, with /æ/ as the second strongest competitor. Thus, overall, the four vowels seemed to fall into two perceptual groups - {/ɑ/, /æ/} and {/ɛ/, /ɪ/}. The two groups correlate with the durations of the vowels, i.e. /ɑ/-/æ/ are long (temporal duration) vowels while /ɛ/-/ɪ/ are short vowels (Fig. 5). Vowel /æ/ was a stronger competitor than /ɑ/ for the short vowels /ɛ/ and /ɪ/ at SNR $\geq$ -16 dB. These confusions are consistent with the second formant frequencies of the vowels (Fig. 5), which would be audible at higher SNRs.

A principal component analysis (PCA) was performed on the 4 $\times$ 4 vowel CM $[P(\mathcal{V}_h|\mathcal{V}_s)]$ to analyze the grouping of vowels. The 4 dimensions of the eigenvectors were rank-ordered from 1 to 4 in the decreasing order of the corresponding eigenvalues. The highest eigenvalue was always unity since the vowel CM was row-normalized and therefore the coordinates along the corresponding dimension (i.e. Dimension 1) were the same for all four vowels [1]. The clustering of the vowels in the 3D eigenspace, when projected on a specific X-Y plane in the eigenspace, is very close to the graph of vowel duration vs. the second formant frequencies. The projection coefficients indicate that dimension X is almost identical to dimension 2, which is associated with the largest eigenvalue of the 3D subspace. Therefore, the vowel duration is the most dominant perceptual cue for vowel discrimination. Addition of masking noise reduces the perceptual distance among the vowels and draws them closer in the eigenspace. The vowel /ɪ/ stays relatively farther from rest of the vowels in presence of noise, which

is consistent with its relatively higher recognition scores, possibly due to its high F_2 [10].

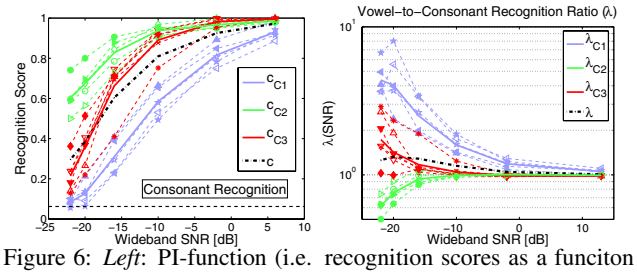

Figure 6: *Left*: PI-function (i.e. recognition scores as a funciton of SNR) for the sixteen consonants. *Right*: Vowel-to-consonant recognition ratio (λ) as a function of SNR, for each consonant. The three solid lines with different darkness represent each of the three consonant sets, while the thick dash-dotted line represent the average values over all consonants.

3.4. Recognition Scores

The speech-weighted noise seems to uniformly mask all the perceptual features (i.e. events) of the vowels. The recognition scores for consonants (Fig. 6, left), however, show much greater variation than the vowels. The curves of consonant scores separate into three groups, which correlate closely with the three consonant sets C1, C2 and C3. The separation of the three sets of curves is more evident in the vowel-to-consonant recognition ratio ($\lambda \equiv \frac{v}{c}$) plots shown in Fig. 6, right panel. The dash-dotted line shows the average value of λ, which was always greater than but very close to unity. Set C1 consonants have λ curves that start rising with addition of small amount of noise, while those for set C3 stay close to unity for wideband SNRs $\geq$-16 dB, but rise sharply below that. However, λ for set C2 is always less than unity and decreases gradually for wideband SNRs $\leq$-10 dB. This shows that recognition scores of some consonants (C2) can be greater than vowels in speech-weighted noise.

The spread of the consonant recognition scores (Fig. 6, left panel) is very large ($\sim$6%-80% at -20 dB SNR) in speech-weighted noise. It is consistent with the consonant scores measured by [11]. In comparison, the spread of the consonant scores in white noise (i.e. MN16-55 data, shown in Fig. 6 of [7]) is very small, nearly half of that observed in speech-weighted noise. Thus, the white noise masks the consonants much more uniformly as compared to the speech-weighted noise. This implies that the events for the consonant sounds are distributed uniformly over the bandwidth of speech. The events important for recognizing the C2 set consonants are at the higher frequencies that are relatively less masked by the speech-weighted noise. The events for the vowels are mostly at the lower frequencies, which are uniformly masked by the speech-weighted noise.

These inferences are consistent with the work of [12], who show that while the vowel recognition is always better than the consonant recognition ($\lambda > 1$) in low-pass filtered condition, the high-pass filtering could make the vowel recognition either smaller or greater than the consonant recognition. The low-pass filtered speech contains almost all vowel events but only few, low-frequency consonant events, thus making the vowel scores always greater than consonant scores. On the other hand, the high-pass filtered speech contains very few or no vowel events. High-pass filter also removes the low-frequency consonant events, resulting in an average consonant score that could be greater or smaller than the vowel score.

4. Conclusions

1 In presence of speech-weighted noise, the CV syllables perceptually group into three sets C1, C2 and C3. These sets are determined by the consonant in the CV and not by the vowel (Fig. 1).

2 Set C1 consonants (/f/, /θ/, /v/, /ð/, /b/) have recognition scores smaller than the vowel scores. The confusion patterns for these consonants (along with that of /m/ from set C1) are influenced by the vowel.

3 Set C2 consonants (/s/, /ʃ/, /z/, /ʒ/, /t/) have recognition scores greater than the vowel scores and confusion patterns of these consonants are not affected by the vowel. Within set C2, /s/-/z/ and /ʃ/-/ʒ/ form highly asymmetric perceptual groups, biased in favor of the voiced consonant in presence of noise (Fig. 2).

4 The consonant confusions among the three consonant sets correlate with the spectral energy in the consonants (Fig. 3).

5 Vowel duration is the most dominant acoustic feature in the perceptual grouping of the vowels (i.e. /ɑ/-/æ/ and /ɛ/-/ɪ/), followed by the second formant frequencies (Fig. 5).

6 A comparison of UIUCs04 results with the past work suggests that the events for vowels are located at low frequencies, while those for consonants are spread uniformly over the entire bandwidth of speech.

5. References

[1] J. B. Allen, *Articulation and Intelligibility* (Morgan and Claypool Publishers, USA, 2005), series editor B. H. Juang.

[2] R. P. Lippman, "Speech recognition by machines and humans," Speech Communication **22**, 1–15 (1997).

[3] M. D. Wang and R. C. Bilger, "Consonant confusions in noise: a study of perceptual features," J. Acoust. Soc. Am. **54**(5), 1248–1266 (1973).

[4] J. Sroka and L. D. Braida, "Human and machine consonant recognition," Speech Comm. **45**(4), 401–423 (2005).

[5] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some english consonants," J. Acoust. Soc. Am. **27**(2), 338–352 (1955).

[6] S. A. Phatak, J. B. Allen, and A. Lovitt, *Quantifying the perceptual quality of speech database* (submitted to INTERSPEECH 2006, Sep 17-21, 2006).

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

[8] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of american english vowels," J. Acoust. Soc. Am. **97**(5), 3099–3111 (1995).

[9] G. E. Peterson and H. L. Barney, "Control methods used in a study of vowels," J. Acoust. Soc. Am. **24**(2), 175–184 (1952).

[10] S. Gordon-Salant, "Some perceptual properties of consonants in multitalker babble," Perception & Psychophysics **38**, 81–90 (1985).

[11] K. W. Grant and B. E. Walden, "Evaluating the articulation index for auditory-visual consonant recognition," J. Acoust. Soc. Am. **100**(4), 2415–2424 (1996).

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