ACCOUNTING FOR PERCEPTUAL IDENTIFICATION OF CONSONANTS AND VOWELS THROUGH ACOUSTIC DISSIMILARITY

Jianxia Xue

Department of Electrical Engineering
University of California at Los Angeles
420 Westwood Plaza,
Los Angeles, CA 90095, USA

ABSTRACT

This study was undertaken to examine relationships between acoustic speech measures and auditory phonetic perception. The hypothesis of the study was that physical dissimilarity of the acoustic measures could substantially account for perception. Speech samples of 22 Consonant-/a/ syllables and 14 /h/-Vowel-/d/ syllables were spoken by two talkers. The stimuli were processed by vocoders with two different filterbanks. Forced-choice perceptual identifications were obtained from 6 normal hearing participants for each talker, vocoder, and syllable set. Confusion data were analyzed using multidimensional scaling, and Euclidean distances among stimulus phonemes were computed. For each pair of stimuli (within the same talker, vocoder, and syllable set), physical Euclidean distances were computed within frequency channels and averaged across time. Multilinear regression was used to transform the Euclidean physical distances to perceptual distances. Evaluation using Pearson r showed that the transformed physical distances correlated with perceptual distances between 0.55 and 0.96 (30% to 91% variance accounted for), depending on the talker, vocoder, and syllable set. The results indicated that the distinctiveness of the speech signals can account for perceptual dissimilarity structure by using a linear transformation.

1. INTRODUCTION

One general theoretical view of speech perception is that auditory cues are projected to segmental linguistic units [1]. Another view is that perceivers recover articulatory gestures from auditory stimuli [2]. Still another view is that auditory information is evaluated in terms of features and then integrated in terms of prototypes [3]. However, none of these proposed mechanisms has yet been adequately confirmed. The current study investigated the possibility that by appropriately transforming physical stimulus measures, the perceptual dissimilarity structure of speech consonants and vowels could be accounted for directly, without recourse to an intervening linguistic level of coding, such as features or gestures [4][5].

2. METHOD

In this study, a custom real time vocoder was used to process acoustic stimuli for perceptual and physical measures. The vocoder (see Fig. 1) effectively degraded the phoneme identification in the perceptual experiment, thus providing sufficient confusions for modeling perceptual dissimilarity. In each channel of the vocoder, the carrier frequency was the center frequency of the band-pass filter. A gain vector was chosen to flatten the ANSI-standard speech-shaped spectrum. The vocoder output was the summation of a set of sinusoids whose amplitudes were modulated by the energy of the corresponding band of original speech and an equalization gain factor [6]. Two different filterbanks were used in this study. The F2 High-pass (F2HP) filterbank had 15 channels with the center frequencies equally spaced from 825 to 2625 Hz and at 3115 Hz and 3565 Hz. The F1 filter bank had 12 channels with the center frequencies ranging from 75 to 900 Hz.

Fig. 1. Logic block diagram of vocoder system

2.1. Participants, material and stimuli recording

There were 4 sets of 6 participants who were assigned to a particular talker and vocoder experiment. They were all native speakers of American English, from 18 to 45 years of age, and with normal hearing.

The speech material comprised two tokens each of 22 C/a/ syllables and 14 /h/-Vowel-/d/ syllables. The 22 American English consonants were /y, w, r, l, m, n, p, t, k, b, d, g, h, ð, ð, θ, s, z, f, v, j, z, tʃ/. The 14 American English vowels were /i, ì, ì, a, a, e, æ, æ, ð, ð, ð, ð, ð, ð, ou, au, o, ò, ʊ, u/. The syllables were recorded by one male and one female talker. The four data sets used for physical measures were digitized from the vocoded stimuli. The original acoustic samples and vocoded acoustic samples were recorded simultaneously on different channels of the sound file. The original acoustic signal provided time-alignment for the vocoded acoustic samples.
2.2. Perceptual testing procedures

Testing was conducted in a sound-treated (IAC) booth. Stimuli were presented over earphones. A simulated keyboard with 22 consonants or 14 vowels was displayed on a monitor. Participants responded by selecting a response using the computer mouse. In each test, participants started with practice block, followed by 2 test blocks. Stimuli were presented in pseudorandom order within blocks. For each syllable, each vocoder and each talker, there were 120 responses (2 tokens x 10 trials x 6 participants). The experiment was completed in about 2 hours for each participant.

2.3. Perception distance analysis

Phoneme confusion matrices. Phoneme identification data were aggregated and formed 8 stimulus-response confusion matrices ([22 consonants or 14 vowels] x [F1 or F2HP vocoder] x two talkers). The confusion matrices listed the percentage of trials for which each response label was chosen.

Multidimensional scaling (MDS). MDS was used to examine the perceptual dissimilarity structures of the consonants and vowels. The 8 phi-square transformed confusion matrices [8] were used as input data to MDS. The analysis generated spatial representations of the consonants and vowels for the F1 and F2HP vocoders and the male and female talkers. Euclidean distances were calculated using the 3-D solutions, resulting in 231 distances between consonant pairs and 91 distances between vowel pairs.

2.4. Physical distance analysis

Analysis window. The acoustic data were all re-sampled to 16 KHz. The onset of the signal corresponded to the onset of the original audio channel. The consonant and the /a/ onsets were labeled for the CV stimuli. For /h/V/d/ stimuli, the vowel onsets and offsets were labeled. The mean consonant durations (including the transition to /a/) were 151 msec and 155 msec for the male and female talkers, respectively. The mean vowel durations were 280 msec and 421 msec for the male and female talkers, respectively. These statistics were used to determine the duration for processing.

For all C/a/ syllables, the analysis window was defined from the consonant onset for a fixed duration forwards. For all /h/V/d/ syllables, the analysis window was defined from the offset of the vowel for a fixed duration backwards. This approach was taken, because the offset of the vowel was less ambiguous than its onset, and because this assured that diphthongs were adequately included in the analysis.

The window over which analyses were conducted was systematically varied in duration and in frequency resolution (see below). The Pearson $r$ was re-calculated for each analysis. Then, the Fisher’s $Z$ transformation [7] and its inverse transformation were used to find confidence intervals for the mean Pearson correlation score.

Feature extraction. The vocoded acoustic data in the time domain were transformed to the frequency domain using the FFT for every 10 msec with a rectangular window of 20 msec. Thus, a matrix was formed for each phoneme with frequency channels in columns and time frames in rows as follows:

$$
F^{S,R} = \begin{bmatrix}
\mathbf{f}_{1,1} & \cdots & \mathbf{f}_{1,Ch} \\
\vdots & \ddots & \vdots \\
\mathbf{f}_{Fr,1} & \cdots & \mathbf{f}_{Fr,Ch}
\end{bmatrix},
$$

where $S$ and $R$ were syllable and token index; $Ch$ and $Fr$ were the number of channels and frames, respectively. The number of channels depended on the vocoder type and frequency resolution. The number of frames depended on the analysis window duration. For example, for a 256-point FFT, the $F$ matrix had 17 channels, spaced 62.5 Hz apart for the F1 vocoder, and 65 channels with the same spacing for the F2HP vocoder. For a 300-msec window, the number of frames was 29.

Physical distance matrices. Physical Euclidean distance matrices were formed as follows for all possible pairs of phonemes (within talker, vocoder, and syllable set), using the $F$ matrix for each phoneme. A vector of physical distance between a pair of syllables was calculated as follows:

$$
d^{S1−S2} = \sqrt{\sum_{R=1}^{Fr} \left( \sum_{f=1}^{Fr} (F_{f}^{S1,R} − F_{f}^{S2,R})^2 \right)},
$$

where $f$ was the frame index, and $d^{S1−S2}$ was the physical distance vector (a row in $D$) between syllables $S1$ and $S2$. Table 1 lists the dimensionality of physical distance $D$ across syllable sets and vocoders.

<table>
<thead>
<tr>
<th>Table 1. Dimensionality of physical Euclidean distance space $D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition</td>
</tr>
<tr>
<td>Dimension</td>
</tr>
</tbody>
</table>

2.5. Estimating perceptual distance from physical distance

Multilinear regression was used to linearly transform physical distances to perceptual distances. The transformation vector $w$ is optimal in the least square sense. $D$ is the matrix of all physical distances, and $b$ is the vector of all perceptual distances. $w$ was computed using Eq. 3.

$$
w = (D^T D)^{-1} D^T b.
$$

The performance of this linear transformation was evaluated using the Pearson correlation $r$ between the transformed physical distances $Dw$ and the perceptual distances $b$.

3. RESULTS

3.1. Overall perception results

The perceptual identification accuracy varied across conditions from 13% to 87% as shown in Table 2. All scores were substantially above chance performance levels (4.3% for consonants and 6.7% for vowels) [8]. For both talkers, the F2HP vocoder was more intelligible than the F1 vocoder, and the vowels were more intelligible than the consonants. Performance varied across a larger range for the female’s speech. Given differences in the spacing of the voice harmonics of the two talkers and the differences in the spacing of the filters across vocoders, it was anticipated that there would be a range of intelligibility scores in this study. The extent to which the varied levels of performance and patterns of confusions could be accounted for using the multilinear regression method can be considered to be a rigorous test of those methods.
Table 2. Perceptual identification accuracy score

<table>
<thead>
<tr>
<th>Talker</th>
<th>F1-C</th>
<th>F2HP-C</th>
<th>F1-V</th>
<th>F2HP-V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>27.4%</td>
<td>72.2%</td>
<td>46.5%</td>
<td>80.6%</td>
</tr>
<tr>
<td>Female</td>
<td>12.7%</td>
<td>70.5%</td>
<td>27.3%</td>
<td>87.6%</td>
</tr>
</tbody>
</table>

3.2. Effect of dimensionality of F matrix

The number of channels and frames in the physical F matrix influenced the magnitude of the obtained Pearson rs. Table 3 summarizes results across all of the analyses on window duration effect for both talkers, both vocoders and both syllable sets. For consonants, analysis window durations from 150 to 400 msec were tested, and for vowels, durations from 200 to 450 msec were tested. The maximum, mean, and standard deviation of Pearson $r$ ($max(r)$, $\bar{r}$ and $STD(r)$) were obtained from 26 window durations in each condition. In Table 3, the upper and lower bound ($r_l$ and $r_u$) of the 95% confidence interval for $\bar{r}$ were obtained with Fisher’s $r$-to-$Z$ transformation and inverse transformation. All of the durations tested were shown not to produce $r$s that differed significantly among themselves.

Frequency resolution had a stronger influence on the results. For example, for the F1 vocoder applied to the male talker’s vowels, the $\bar{r}$ varied significantly between .86 and .96 for 31.25 and 15.625 Hz resolution, respectively. The usable frequency resolution was limited by the row size of the $F$ matrix (e.g., under the F2HP-V condition, singular matrix inversion occurred in multilinear regression at 31.25 Hz). In this study, 62.5 Hz resolution was selected for all the analyses.

Table 3. Statistics and confidence intervals for Pearson $r$

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>F1-C</th>
<th>F2HP-C</th>
<th>F1-V</th>
<th>F2HP-V</th>
</tr>
</thead>
<tbody>
<tr>
<td>max($r$)</td>
<td>0.55</td>
<td>0.70</td>
<td>0.84</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>$\bar{r}$</td>
<td>0.53</td>
<td>0.66</td>
<td>0.83</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>$STD(r)$</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>$r_l$</td>
<td>0.61</td>
<td>0.72</td>
<td>0.88</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>$r_u$</td>
<td>0.43</td>
<td>0.58</td>
<td>0.76</td>
<td>0.80</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>F1-C</th>
<th>F2HP-C</th>
<th>F1-V</th>
<th>F2HP-V</th>
</tr>
</thead>
<tbody>
<tr>
<td>max($r$)</td>
<td>0.82</td>
<td>0.67</td>
<td>0.76</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>$\bar{r}$</td>
<td>0.78</td>
<td>0.62</td>
<td>0.74</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>$STD(r)$</td>
<td>0.05</td>
<td>0.06</td>
<td>0.02</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>$r_l$</td>
<td>0.83</td>
<td>0.70</td>
<td>0.82</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>$r_u$</td>
<td>0.72</td>
<td>0.54</td>
<td>0.62</td>
<td>0.88</td>
<td></td>
</tr>
</tbody>
</table>

3.3. Estimation from physical to perceptual distance

Fig. 2 presents scatter plots of perceptual distance versus estimated perceptual distance (i.e., transformed physical distance) for the $max(r)$ results for each condition (see also Table 3). Across the eight sets of results in Fig. 2, six of the $r^2$ values represented between 50% and 91% variance accounted for. That is, the transformation was substantially successful. Two of the sets of results represented less than 50% of the variance, 30% F1 (for the male consonants) and 45% F2HP (for the female consonants).

4. DISCUSSION

The Pearson $r$ evaluation measure showed that a linear transformation of physical acoustic dissimilarity could be highly correlated with the auditory perceptual dissimilarity measures. This correlation was obtained without any additional linguistic coding of the stimuli such as one involving features or articulatory gestures. This supports the view that the physical signal dissimilarity can account directly for the segmental distinctiveness of consonants and vowels in English.

Comparison across Tables 2 and 3 showed that the degree of this correlation was not straightforwardly related to perceptual intelligibility. The explanation for this could be that perceptual intelligibility scores represented only the results on the diagonal of the confusion matrices, whereas the linear transformation took into account the pattern of responses across the entire matrix. Even a low intelligibility condition could result in good performance of the transformation, when the confusions in the matrix are systematic (e.g., the F1-C condition for the female talker).

As suggested above, differences in patterns of confusions were anticipated across talkers, vocoders, and syllable sets. That the results were very good for six of the confusion matrices suggests that the method is quite robust in capturing the relationship between perceptual and physical distances. The least successful analysis involved the F1 vocoder, male talker, and consonants. Examination of Fig. 2b suggests that the relatively low $r$ was the result of a restricted range along the perceptual distance dimension. Examination of Fig. 2b suggests that the relatively low $r$ for the female talker’s F2HP data might also be due to restricted physical distances, consistent with the results in Fig. 2a.

The physical dissimilarity measurement was based on the FFT directly. No attempt was made to model effects due to the human auditory periphery. The male talker’s low fundamental frequency likely resulted in more information for the physical distance measures than was processed perceptually. Even though the transformation based on the raw FFT measurement resulted in an average of 61% variance account for across conditions, it is likely that to improve the results, the physical stimuli need to be transformed non-linearly to match the non-linearities of the auditory system. Furthermore, it is likely that at the level of perceptual processing there are additional non-linearities.

5. CONCLUSIONS

In this paper, we examined the relationship between auditory perceptual dissimilarity and physical dissimilarity measures for consonant and vowel nonsense syllables. We showed that a linear transformation from physical dissimilarity to perceptual dissimilarity could account for up to 91% of the variance in the data. In general, the linear model performed better for vowels (an average of 74% variance accounted for) than consonants (an average of 48% variance accounted for). In general, the analysis window length was found to be not particularly sensitive over quite a large range. These results support the view that the perceptual structure of English consonants and vowels can be accounted for by an appropriate linear transformation of the physical stimulus structure, and that this can be done across talkers, filter configurations, and syllable sets. Furthermore, these results support the hypothesis that speech perception operates directly on the physical acoustic structure of the consonant and vowel segments in English.
6. ACKNOWLEDGMENT

This research was supported by NIH/NIDCD 00695. We thank John Jordan and Brian Chaney for technical assistance and Dr. Edward Auer for commenting on the research.

7. REFERENCES