ACOUSTIC CUES TO PERCEPTION OF VOWEL QUALITY

Fernández, S., Feijóo, S.
Departamento de Física Aplicada
Universidad de Santiago de Compostela
15706 Santiago de Compostela, SPAIN
E-mail: fasanti@usc.es

ABSTRACT

Perceptual and acoustic spaces obtained using MDS and PCA, respectively, for a set of Spanish vowels uttered in various fricative contexts are compared. Acoustic cues to vowel quality were studied through examination of the weights given by PCA to specific spectral regions. The acoustic space formed by the first two formant frequencies was also evaluated. The results show that MDS has serious difficulties to be useful for the study of Spanish vowels. PCA offers a gross characterization compared to formant frequencies. Neither of them completely explains the perception of vowel quality.

1. INTRODUCTION

There has been a great deal of research concerning the acoustic correlates of vowel quality. The first and more evident choice are the resonant frequencies of the vocal tract, which provide an easy interpretation of vowel perception: the vocal tract configuration at a given moment is characterized as an acoustic (resonant) tube, which is defined by a series of formant frequencies. The locations of these formants (and maybe their amplitude) are responsible for the perceived quality of the vowel.

The main advantage offered by the acoustic characterization of vowels through formants is its compactness: Two or three values (F1, F2, F3) are enough. The two-formant F1/F2 space has the additional advantage of being an easy and straightforward spatial representation. Nevertheless, despite these advantages, the formant concept has some drawbacks. First, from the point of view of the acoustic characterization of vowels, it has rarely been used, for instance, in ASR technology. There are three main reasons for this: the problem of formant definition at even moderately high fundamental frequencies, the merging of formants, and the interaction with antiformants. As a consequence, automatic determination of formants is highly problematic. An additional reason is the variation of formant values with both sex and consonantal context: the coarticulatory influence of the consonant on the vowel in CV syllables alters the location of the resonant frequencies of the vowel in a way that augments the already considerable overlapping among vowel categories in the formant space. Second, from the point of view of vowel quality perception, the compactness of the representation implies that some auditorily relevant information may have been discarded. That means that formants may not be the sole determinants of perceived vowel quality (Hillenbrand & Nearey, 1999). Moreover, the aforementioned influence of consonantal context on formants may cause additional problems (Fernández et al., 1998).

The mentioned problems have provoked a quest for other correlates of vowel quality. However, it is not clear that alternative representations can solve all those problems, and some may introduce additional drawbacks, like a much higher dimensional order (Zahorian & Jagharghi, 1993; Hillenbrand & Nearey, 1999). It would be desirable, then, to have a compact acoustic representation of vowels that could be computed automatically and with ease. One such a representation may be based on Principal Components Analysis (PCA). PCA involves forming linear combinations of a set of variables. In that way a n-dimensional acoustic space can be transformed into a two or three-dimensional space. The new variables (principal components) are then just a transformation of the original ones, with the additional advantage that the explained variance of the sample for each principal component is now available. Moreover, once the original set of variables has been determined, the new, reduced space, can be automatically computed. Then, it is critical to choose a set of variables that offer a good description of vowel quality. A set of mel-frequency scaled filter outputs could offer a good description of the vowel spectra in which both local and gross spectral information are contained. The combination of PCA with mel-scale filter outputs offers another important feature. Since the principal components are linear combinations of the old variables, a series of factor loadings reflecting the particular contribution of each old variable to the new variables is obtained. In this way, the relevant characteristics of the vowel spectrum that most contribute to vowel quality can be detected by examination of the factor loadings corresponding to each new variable, i.e. how each spectral region contribute to vowel quality.

Multidimensional scaling (MDS) analysis offers the possibility of obtaining a spatial description of listeners’ perception. Several authors (see for instance Pols et al., 1969) have exploited the idea of comparing PCA- and MDS-derived spaces. Nevertheless, they only included /hVd/ tokens. Besides, they did not make a thorough analysis of factor loadings.

In a previous paper (Fernández et al., 1998), the perception of vowels in the context of the voiceless fricatives (/F, s, j, x/, and the affricate /F/) was studied. The perceptual results revealed that perception of vowel quality was influenced by the presence/absence of the consonantal cue: the vowel, per se, did not seem to offer all the required information for vowel quality perception. A PCA carried out on a set of 18 mel-frequency scaled filter outputs showed that there was a considerable overlap among vowels, and marked differences among fricative contexts.

In this paper the acoustic space obtained through PCA is compared with the perceptual space of the same dataset obtained...
through MDS. MDS extracts the dimensions relevant to the perception of vowel quality. A bidimensional MDS, thus, offers a visual perceptual space similar to that obtained through a bidimensional PCA. Both spaces can be compared between them, and can also be compared with the bidimensional F1/F2 space. If the agreement between MDS and PCA spaces is good, then inspection of the PCA’s factor loadings will allow us to know which spectral regions contribute to vowel perception. That information can contribute to shed light on the actual role played by formants in the definition of vowel quality, if a good correlation between the PCA space and the F1/F2 space is obtained.

2. MATERIALS

The vocalic segments included in this paper are the same studied in Fernández et al. (1998) as part of a larger database. Two-syllable Spanish words whose first syllable was formed by the combination of the fricatives /f, o, s, f/ and /s/ and the affricate /f/ with the five Spanish vowels /a, e, i, o, u/ were uttered by 5 male native speakers of Spanish (Galician region). The total number of tokens is 150 = 5 vowels × 6 fricative contexts × 5 male speakers.

The tokens were sampled at 20 kHz with 12 bits of precision, and band pass filtered with cutoff frequencies of 100 Hz and 9.2 kHz. The point previous to the first clear vocalic pulse in the first (FV) syllable of every word was marked, and the following 51.2 ms of vowel were isolated. These stimuli include the transition of the fricative into the vowel and part of the vocalic nucleus, but they do not include the transition out of the vowel. These 150 vocalic segments form the corpus for this study.

3. PERCEPTUAL SPACE

Perceptual experiments with these stimuli were carried out in Fernández et al. (1998). 12 subjects, all native speakers of Spanish, served as listeners. The perceptual experiment consist of identifying the vowel corresponding to each one of the 150 vocalic segments. Stimuli were presented in random order. Listeners could carry the experiment out at their own pace. Only one repetition was allowed, after which the listener had to choose one of the possible options: the Spanish vowels /a, e, i, o/ and /h/, or the option another sound. The obtained confusion matrix can be seen in table 1. Overall identification rate was 80.3%. It was always around 80% for each vocalic category except /h/, which attained a lower identification rate: 73.9%. Differences among fricative contexts showed up. Identification rates were 81.3%, 85.0%, 88.3%, 68.7% and 86.3% for vowels pronounced in the context of /f, l, r, s, f, f/ and /h/, respectively. Therefore, vowels were worse identified in the context of /f/ and /h/ than in any other fricative context. This effect was especially marked for the vowel /l/: only 53.3% correct identification rate in the context of /f/ and 60.0% in the context of /h/. Detailed analyses of variance of the results are available in Fernández et al. (1998).

Multidimensional scaling analysis was carried out in order to depict the listeners’ perceptual space. Since differences among fricative contexts were notorious, an individual differences (INDSCAL) model was suggested. For this model, the confusion matrices, separately for each fricative context, are entered into the analysis. Thus, general coordinates for the stimuli —the five vowels in this particular case—are obtained along with different weights (one for each dimension of the perceptual space) for each fricative context. These weights indicate how different the perceptual space for a particular fricative context is with respect to the general perceptual space. The perceptual space for a particular context can be obtained by multiplying, in each dimension, its assigned weight by the general coordinates of the stimuli. Confusion matrices show usually asymmetry and bias. An asymmetric individual differences (ASYNDSCAL) model was, therefore, chosen. The alternating least squares scaling (ALSCAL) computer program was used to carry out this type of analysis. Three- and two-dimensional perceptual spaces were obtained but three-dimensional solutions never achieved significantly higher correlations with listeners’ confusion matrices than two-dimensional spaces (correlation for two-dimensional spaces was very high). Besides, their interpretability was always more difficult. This corroborates Fox et al.’s (1995) finding stating that Spanish listeners use two perceptual dimensions for identifying vowels while English listeners use three; although their study only included the English vowels /a, e/ and /l/. They also found that perceptual dimensions did not agree with formant frequencies’ space. The resulting two-dimensional perceptual space obtained a high correlation ($r^2 = 0.96$) with listeners’ judgments. Nevertheless, it hardly resembled listeners’ perception. The stimuli were located in the vertices of a triangle: /a/ in one vertex, /e/ and /l/ in another vertex very close to each other, and /o/ and /h/ slightly separated in the third vertex. The MDS space did not explain either the confusions among vocalic categories or the effect of the fricative context. A possible explanation of this result could be that there is little information about the perceptual distances among vocalic categories in the confusion matrices, and also little difference among the confusion matrices for each fricative context. The split-half method was applied in order to assess the validity of the perceptual space. Different solutions were obtained for both halves: one similar to the aforementioned space ($r^2 = 0.98$), and the other one more similar to the vocalic triangle ($r^2 = 0.94$). The results were not conclusive since the correlation between responses profiles for both halves (6 listeners were included in each half) were not very high ($r^2 = 0.5$), nevertheless the results suggest that various different perceptual spaces can represent the confusion matrices’ data. Therefore the analysis along with the perceptual experiment does not seem to be adequate to study the perception of Spanish vowels. Since one of the possible perceptual spaces explaining listeners’ responses could be the well-known vocalic triangle, the initial configuration of the MDS analysis was modified in order to resemble a neutral triangle. The perceptual space is showed in figure 1. The /a, e, i, o/ and /h/ labels indicate which cluster corresponds to each vocalic category. In each cluster, the location of each fricative context is showed. Initial stimuli configuration given to MDS analysis is also indicated in the figure.

<table>
<thead>
<tr>
<th>/a/</th>
<th>/e/</th>
<th>/h/</th>
<th>/o/</th>
<th>/u/</th>
<th>another</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>85.6</td>
<td>2.5</td>
<td>0.0</td>
<td>6.9</td>
<td>0.0</td>
</tr>
<tr>
<td>/e/</td>
<td>2.2</td>
<td>80.4</td>
<td>9.4</td>
<td>2.2</td>
<td>2.2</td>
</tr>
<tr>
<td>/h/</td>
<td>0.0</td>
<td>21.7</td>
<td>77.8</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/o/</td>
<td>5.0</td>
<td>2.2</td>
<td>0.0</td>
<td>73.9</td>
<td>13.1</td>
</tr>
<tr>
<td>/u/</td>
<td>0.0</td>
<td>0.8</td>
<td>3.1</td>
<td>7.8</td>
<td>83.6</td>
</tr>
</tbody>
</table>

Table 1: Confusion matrix resulting from the identification by 12 listeners of Spanish vowels pronounced in fricative context. Overall identification rate was 80.3%. Percent correct identification was 81.3%, 85.0%, 88.3%, 71.7%, 68.7% and 86.3% in the context of /f, l, r, s, f, f/ and /h/, respectively.
The correlation of the perceptual space with information in listeners’ confusion matrices was $r^2 = 0.96$, and it was also very high for every fricative context: $r^2 = 0.91$ for /f/, $r^2 = 0.93$ for /sl/, $r^2 = 0.99$ for /sl/, $r^2 = 0.98$ for /fl/, $r^2 = 0.96$ for /fl/, and $r^2 = 0.96$ for /sl/. The results show little difference with respect to the initial configuration, except in the context of /sl/. The perceptual space does not show more confusion in the context of /sl/ and /fl/, and for the vowel /sl/, as expected.

4. ACOUSTIC SPACE

4.1. Principal Components Analysis

The method used to carry out the principal components analysis (PCA) can be seen in Fernández et al. (1998). In order to have a significant number of tokens five additional male speakers were included. The number of tokens included in the PCA was 300 = 5 vowels × 6 fricative contexts × 10 male speakers. Results are presented for the same five male speakers included in the perceptual experiments. The spectra of two 25.6 ms windows covering the 300 vocalic segments of 51.2 ms were computed. A set of 18 mel-scaled triangular filter outputs integrated the spectral information from 200 Hz to 4.5 kHz. The level in decibels of every filter output was calculated. The set of 18 filter outputs for each window of 25.6 ms was level-normalized: the average level of all 18 bands was subtracted from the separate band levels. The resulting filter outputs for each window were also averaged. Now, 18 filter outputs represent the 51.2 ms wide vocalic segment. Finally, speaker normalization was carried out: for each speaker and fricative context, the average of all five vowels calculated spectra was subtracted from the five particular spectra. PCA was applied to the set of 18 filter outputs that represent the spectral information in the 51.2 ms wide vocalic segments of the 300 tokens. Two principal components were extracted. The first principal component explained 41.9% of the variance, while the second one explained 15.3% of the variance in the sample. The third principal component only explained 6.7% of the variance. The two-dimensional space was VARIMAX rotated in order to enhance interpretability of the factor loadings. Factor loadings are the correlations between the variables and the principal components. Thus, a high squared factor loading indicates that the variable is well represented by the principal component. Besides, a high value of a principal component indicates that the values of the variables with high squared factor loadings for that principal component are also high. Similarly, if the value of a principal component is low this indicates that the values of the variables with high squared factor loadings for that principal component are low. The inspection of factor loadings will allow us to relate the acoustic space to the spectral information. The acoustic space is shown in figure 2. Again, /i/, /e/, /o/ and /u/ labels indicate which cluster corresponds to each vocalic category; and, in each cluster, the mean location of the 5 male speakers’ tokens for each fricative context is shown. As can be seen, vocalic categories are overlapped to a large extent (see also Fernández et al., 1998). /sl/ does not seem to be more confusing than other categories, such as /sl/ or /sl/. Vowels in the context of /fl/ and /fl/ show a smaller acoustic space than in any other context, which is in agreement with listeners’ responses, although a less confusing vocalic space could be expected since identification rates were not so low in those fricative contexts. Overall correlation of the acoustic space with the perceptual space was ($r^2 = 0.63$). Factor loadings for both principal components are plotted in figure 3. The first principal component explained 41.9% of the variance. It weights the region between 500 and 1500 Hz with negative value, and gives positive weight to the region above 1700 Hz. This seems to correspond to a second formant distinction but obtained with a gross characterization (see figure 2). The second principal component explained 15.3% of the variance and although it could be say that it corresponds to the first formant (see figure 2), its corresponding factor loadings indicate that it is a combination of the first formant (peaks around 700 Hz) and the second formant (peaks around 1500 Hz).

The results did not completely agree with listeners’ responses. Anyway, since the obtained perceptual space is in doubt, we can not accurately conclude if PCA extracts acoustic information relevant to the perceptual identification of vowel quality to a greater or lesser extent.
4.2. Formant Frequencies

The acoustic information extracted by PCA did not completely agree with formant frequencies. In order to shed some light on the role played by formant frequencies in the identification of vowels, the acoustic space formed by F1 and F2 was obtained. The 150 tokens were resampled at 10 kHz. LPC spectra was computed over several windows of 15 ms, overlapped 10 ms, covering the whole vocalic segment. The poles of every LPC spectra were extracted. The resulting trajectory of the formant frequencies along the vowel was examined and the first and second formant frequencies were manually obtained from the steady-state region of that trajectory, approximately in the fifth vocalic pulse. That region was always located in the first 51.2 ms. The acoustic space formed by the first two formant frequencies can be seen in figure 4. /a, e, i, o/ and /u/ labels indicate which cluster corresponds to each vocalic category. In each cluster, the location for each fricative context pooled over 5 male speakers is showed. The acoustic space shows extremely well separated vocalic categories and almost non-existent contextual effects. Correlation of this space with the perceptual space was very high: \( r^2 = 0.81 \) for the dimension corresponding to F1 and \( r^2 = 0.85 \) for the dimension corresponding to F2. It was always above 0.90 for every fricative context separately. Correlation with PCA space was \( r^2 = 0.61 \) for the dimension corresponding to F1 and \( r^2 = 0.90 \) for the dimension corresponding to F2. Since the absence of contextual effects could be a consequence of studying the formant frequencies in the vocalic nucleus, formant frequencies were also manually extracted in the beginning of the vowel, approximately, in the first pulse. The only contextual effect was that vocalic triangles of /ʃ, ʃʃ/ and /θ/ to a lesser extent, were smaller than in the other contexts and were clearly displaced towards the /θ/ vowel. Authors could verify that clear, but short, diphthongs can be perceived in those contexts. Nevertheless, this does not completely explain listeners' perception regarding contextual effects.

5. CONCLUSION

Although several studies on Dutch vowels (see, for instance, Pols et al., 1969) found MDS useful, our results do not state the same for Spanish vowels. In Spanish only five vowels are available and this can limit the performance of MDS. Besides, identification experiments can not be suitable for the task. Similarities experiments might result in richer matrices that give more information to MDS regarding perceptual distances among stimuli. As regards the acoustic space, PCA using mel-scaled filter outputs has difficulties for extracting acoustic information when contextual effects are included in the sample. Previous studies on vowels using PCA only included /hVd/ tokens (see for example, Pols et al., 1969). The formant-like space obtained with PCA did not completely explain either listeners' perception of vowels or the associated contextual effects. The acoustic space provided by the first two formant frequencies is not enough to explain those effects either. Although formant frequencies clearly distinguish among vocalic categories, their relevance for the definition of vowel quality is still in doubt. Further analyses on the distributions of vocalic categories on the acoustic and perceptual spaces are needed in order to confirm this point.

6. REFERENCES


