New Signal Features for Robust Identification of Isolated Vowels

Aníbal J. S. Ferreira

University of Porto / Seegnal Research, Portugal

a.j.ferreira@ieee.org

Abstract

Current signal processing techniques do not match the astonishing ability of the Human Auditory System in recognizing isolated vowels, particularly in the case of female or child speech. As didactic and clinical interactive applications are needed using sound as the main medium of interaction, new signal features must be used that capture important perceptual cues more effectively than popular features such as formants. In this paper we propose the new concept of Perceptual Spectral Cluster (PSC) and describe its implementation. Test results are presented for child and adult speech, and indicate that features elicited by the PSC concept permit reliable and robust identification of vowels, even at high pitches.

1. Introduction

As a specific case of automatic speech recognition, the automatic recognition of isolated phonemes, namely vowels, is extremely important in real-time interactive applications used to assist in didactic contexts such as language learning, and also in clinical contexts such as speech therapy and rehabilitation. In both areas, children represent the population segment that are most likely to benefit from such interactive applications given their natural and strong motivation to interact with computer-based games offering challenges, competition and discovery, in an appealing way, and using sound as the preferred medium of interaction. However this scenario is not yet the reality since vowel recognition in children has not received much research effort in recent years [1], and the existing solutions addressing this problem show little efficiency and robustness.

The majority of the approaches used to recognize isolated vowels are strongly based on formant analysis using Linear Prediction (LP) techniques. The basic assumption when using LP is that the production of voiced sounds by the human phonetic system can be modeled as an all-pole filter that is excited by a periodic train of pulses. The repetition rate of these (glottal) pulses corresponds to the fundamental frequency of the voiced sound (F0), or pitch, and the poles of the all-pole filter correspond to resonances of the vocal tract, or formants. The spectral positions of the first three formants (F1, F2, F3) are usually considered as acoustic correlates of a given vowel, and are typically estimated using LP [2].

In particular, in the case of male adult speech, and for low sampling frequencies (e.g., 8 kHz), the representation in a plane of the frequencies of the first two formants estimated using LP, allows a clear discrimination between the different vowels. This plane is normally known as the F1-F2 plane. However, the technique is not speaker independent and the discrimination ability of this approach fails in the case of children speech or singing voice [3].

In this paper we study features of speech signals and associated estimation techniques, that are alternative to formant tracking using LP techniques, but that are perceptually reliable and robust in identifying isolated vowels, in a speaker independent way. In section 2 we detail the context and motivation of our research. In section 3 we address previous research results that are relevant for our research. In section 4 we describe a set of experiments that were conducted in order to conclude on the perceptual pertinence of new features able to discriminate vowels and their performance is evaluated and discussed in section 5. Section 6 presents the main conclusions of this paper and suggests possible directions for future research.

2. CONTEXT AND MOTIVATION

Our interest in real-time robust vowel identification emerged when we were involved in a project aiming at building a set of interactive demonstrators on the subject of sound and the human auditory system. One such specific demonstrator was a real time audio/speech analyzer featuring time and frequency visual displays, accurate pitch estimation, accurate identification and visual representation of harmonic structures, and visually-oriented vowel recognition using LP formant tracking; all in real time\(^1\). After fine-tuning the operation of the vowel recognizer for the main Portuguese vowels, we were quite disappointed verifying that LP formant tracking was very reliable and robust for male speakers, but not for female or child speakers, who were our main target audience.

We were familiar with typical problems of LP formant estimation (namely its performance dependency on such factors as order of the LP model, sampling frequency, pitch), and we were also aware of typical solutions used to attenuate these problems (including decimation, pre-emphasis, shifting the poles towards the unit circle, and angular pole separation using frequency warping).

Our previous experience and a review of the literature convinced us that even these solutions do not provide the desired performance in robust identification of isolated vowels. As a consequence, we ought to look for alternative features able to provide high discrimination capability among the different vowels, in a speaker independent way.

Since we are only interested in natural speech, we have made recordings (32 kHz sampling and 16 bit/sample) of the speech sounds corresponding to the most common Portuguese vowels: /a/, /e/, /i/, /o/, /u/. Seven child speakers and one adult speaker have participated in our recordings. For illustration purposes, we will consider in this paper only the recordings by 3 speakers and their performance is evaluated and discussed in section 5. The spectrogram corresponding to FEM2 is displayed in Fig. 1. The recordings are available on the Web page http://www.fe.up.pt/~ajf/aeiou/ that complements this paper.

\(^1\)This PC application can be downloaded from the Web site: http://www.inescporto.pt/cienciaviva/


3. PREVIOUS RESEARCH

Formant estimation using LP techniques faces a fundamental difficulty when analyzing child speech or when analyzing singing signals [3]. In fact, given that the pitch (F0) in these cases (in the order of 300 Hz or even more) is comparable to the frequency of the first formant (F1), the LP-based formant estimation 'locks' instead to the frequencies of the fundamental or partials of the harmonic structure of the vowel. Thus, LP-based formant estimation techniques can be successful only when the spectral envelope is well sampled (i.e., 'illuminated') by the pitch harmonics, i.e., when F0 is significantly lower than the spectral peaks due to formants. Fig. 2 illustrates the F1-F2 vowel scattergram corresponding to subject FEM1. This scattergram has been estimated using an 18th-order LP analysis. It is clear that the high F0 makes it difficult for the LP analysis to correctly estimate the low F1 and F2 formant frequencies that are typical of vowels /u/ and /o/.

Several vowel identification techniques that are alternative to LP-based formant tracking have been proposed in the literature, including spectral analysis on a critical band rate and Gaussian classification [1], Principal Component Analysis using harmonic amplitudes [3], and critical band analysis followed by Perceptual Linear Prediction [4]. Most results show improvements over LP-based formant estimation, however performance for high-pitched vowels is always inferior to that of low-pitched vowels, and frequently results are assessed using synthesized vowels.

4. STUDY OF NEW FEATURES

Our attempts to find alternative signal features that are both stable and perceptually meaningful, led us to extensive experimentation and to the following main conclusions:

1. Perceptually appropriate features in vowel recognition (other than sharp spectral peaks) are linked to a concept that we designate here as Perceptual Spectral Cluster (PSC) and that is broader than the concept of formant. A PSC denotes a spectral region with a significant local concentration of spectral power, the characteristics of which determine important perceptual cues used in vowel recognition. A PSC may be primarily characterized by its center, power, and a measure of shape such as bandwidth or level differences among close harmonic partials,

2. To a great extent, PSC is inherently insensitive to the pitch frequency,

3. The real cepstrum is an appropriate spectral envelope estimation tool in the sense that, contrary to the LP technique, it models equally well spectral peaks and spectral valleys, and therefore it allows good estimation of the boundaries of local PSCs across the whole frequency range.

4. The center frequencies of the first two PSCs can be very reliably computed for each vowel and exhibit high discrimination capability,

5. The power difference between the first two PSCs represents a robust measure of spectral timbre (or sound color) allowing to discriminate among the different vowels.

Given the previous conclusions, and in order to study each feature individually, we prepared an analysis/synthesis simulation environment that included the following processing steps:

1. Spectral analysis using a 1024-point FFT, 50% overlap, and windowing with the square root of the Hanning window [5],

2. LPC analysis using a 18th-order all-pole filter and formant estimation using the frequencies of spectral resonances (only as a reference for test comparison),

3. Smooth spectral envelope estimation by performing the following processing steps: short-time power spectral density (PSD) estimation through the FFT, logarithm of the PSD, IFFT delivering the real cepstrum, short-pass filtering the real cepstrum to 18 coefficients, FFT, inverse logarithm (exponentiation),
4. identification of the boundaries of the first three PSCs by finding maxima and minima on the cepstrum-based smooth spectral envelope model,
5. identification of the center of the first two PSC using a simple center-of-mass like computation,
6. computation of the spectral power of each PSC (as a simple sum of the squared magnitudes of the FFT lines within each PSC boundaries).

4.1. Feature: center frequency of PSC

The above analysis/synthesis environment allowed us to modify sounds and to re-synthesize them. After extensive manipulation of speech signals, we reached the conclusion that from a perceptual point of view, the close proximity of the first two formants (F1, F2) that is typical of back vowels, is perceived a single acoustic entity, and not as two acoustic entities. Significantly, after we draw this conclusion, we were surprised finding that the same perspective had already been presented in the literature under the name of Center of Gravity (COG) [6]. The underlying concept is that the Human Auditory System (HAS) performs a spectral integration spanning a frequency range of about 3.5 Bark. Stimuli with formants closer than this limit are found to be perceptually equivalent to one peak stimuli with the peak position determined by the center gravity of the original two peaks. However, a fundamental difference exists between the COG concept and our PSC concept. While the former is itself a feature consisting in the center-of-gravity, the latter is broader and encompasses opportunities for the extraction of several perceptually relevant features such as the center frequency (equivalent to the center-of-gravity), the bandwidth, the power difference between adjacent harmonic partials, and phase relation between harmonic partials. The main advantage of the PSC concept lies in the fact that it allows for different features being differently weighted in the vowel recognition task according to the characteristics of the speech (e.g., as a function of F0), which is an hypothesis supported by several authors [7, page 3506] regarding the way the HAS processes sound.

It has been concluded that the center frequency of PSCs consists in a fairly stable and robust feature, notably in the case of the first PSC. This feature alone allowed to obtain F1-F2 like plane representations of the center frequencies pertaining to the first two PSCs, with clear separation among all vowels, although depending on the speaker a slightly shift of clusters of symbols could be observed. The re-synthesis of vowel sounds by manipulating the center frequencies of the first two PSC, revealed that the perception of a specific vowel depends much more strongly on the center frequency of the first PSC than on the center frequency of the second PSC.

4.2. Feature: bandwidth of PSC

The bandwidth of the first PSC varies significantly among the different vowels and thus may reveal a significant discrimination capability. Considering this possibility, we have estimated the bandwidth of the first PSC using the -3 db level of the smooth spectral envelope model, relative to its local maximum. Results were rather poor and have indicated that this procedure was not the most appropriate one because at high pitch, the first PSC may include only two or three significant partials of the harmonic structure of the vowel sound, and the resulting envelope model does not reflect the real distribution or spectral power among the different partials.

4.3. Feature: Level difference between the first two PSC

Simple experiments led us to the conclusion that the dB difference between the spectral power pertaining to the first and second PSC, exhibits a high perceptual impact on the vowel discrimination. For example, the vowels /e/ and /i/ are clearly different from /o/ and /u/, respectively, because of this feature. As a first attempt, we have looked at the dB difference between the two local maxima of the smooth spectral envelope model and pertaining to the first two PSC. Because of inaccuracies of this modeling for some signals, its was found that a more reliable measure is the dB difference between the spectral power falling within the first PSC, and the spectral power falling within the second PSC. For all speakers, quite similar results were obtained indicating that the ratio between the spectral power of the first PSC and the spectral power of the second PSC, is much stronger in the case of (back) vowels /a/, /o/ and /u/, than in the case of (front) vowels /e/ and /i/.

4.4. Best discriminating features

As a result of intensive feature experimentation and evaluation, we have concluded that the two best discriminating features are center frequency of the first PSC, and ratio of the spectral power pertaining to the first two PSCs. An illustrative map relating dB-Hz is represented in Fig. 3 regarding speaker FEM1. It can be seen that the vowel symbols are reasonably well clustered and clearly discriminated, contrarily to the observation made for the same speaker using F1-F2 formant scatter plots, as illustrated in Fig. 2. In particular it has been observed that the scattergrams corresponding to the two young female speakers are quite coherent regarding specifically the center frequency of the first PSC. Another very interesting observation was that the ratio between the center frequency of the first PSC obtained for the young female speakers, and that obtained for the male speaker, is approximately 4/3 for all vowels. This suggests that the human recognition of vowels takes into consideration, in an adaptive way, the effect of the pitch frequency on the distribution of the spectral power in frequency. A robust vowel recognizer should therefore reflect this adaptive behavior.

Figure 3: Map of Power Ratio - Center Frequency of PSC for FEM1.
5. DISCUSSION

In order to better assess the vowel discrimination ability of the best PSC features (dB-Hz) relative to the LP formant features (Hz-Hz), a vowel separation measure (VSM) has been evaluated for each pair (i, j) of vowels. The VSM distance between vowel i and vowel j, is defined as:

\[ MS(i,j) = \frac{\text{dist}(m_i, m_j)}{\sigma_i + \sigma_j}, \quad i, j : 1, 2, ..., 5 \]  \hspace{1cm} (1)

where \( \text{dist}(m_i, m_j) \) is the Euclidean distance between the center \( (m_i) \) of vowel cluster i and the center \( (m_j) \) of vowel cluster j, and \( \sigma_i \) and \( \sigma_j \) represent the variance of vowel cluster i and j, respectively. A VSM value higher than 1.0 denotes a good separation between vowel clusters while a VSM value lower that 1.0 indicates a poor separation.

The axis pertaining to the chosen PSC features or pertaining to the F1-F2 formant features have been normalized so as to make it possible to compute distances on a common feature reference plane. After normalization, a single plane has been obtained for each set of features (PSC based or formant based) by overlapping the vowel clusters pertaining to all three speakers. The resulting VSM matrix regarding the chosen PSC features is represented in Table 2 and the VSM matrix regarding the F1-F2 formant features is represented in Table 3. It can be concluded that in the case of PSC features, the smallest separation is observed for the vowel pair /ə/-/ɪ/, which is mainly related to the fact that the spectral power dB difference between the first two PSCs does not provide good discrimination for these two particular vowels. On the other hand, the largest separation is observed for the vowel pair /ɛ/-/ə/ because the spectral power dB difference between the first two PSCs is highest for these two particular vowels. In the case of the F1-F2 formant features, the smallest separation is obtained for the vowel pair /ɪ/-/ʊ/ because in this case the LPC model has an inherent difficulty in correctly identifying the low F1-F2 formant frequencies of vowel /ʊ/ when the pitch is high. The largest separation is observed for the vowel pair /ɛ/-/ɪ/ because in this case the separation between the two F1-F2 formants is the largest and consequently this estimation by the LPC model is more reliable.

Overall, it can be concluded that the chosen PSC features allow a better discrimination among all vowels than the F1-F2 formant features, even at high pitch, because the VSM measure, as shown in Table 2, is consistently higher than 1.0 for all pairs of vowels.

6. CONCLUSION

Our research confirms the idea that the HAS relies on signal features or cues allowing robust recognition of vowels and that are not captured by known techniques. In order to investigate new features providing good vowel discrimination capability, independently of the speaker and notably of its pitch, a significant number of experiments was carried out using time domain techniques, frequency domain techniques, and test procedures involving analysis and synthesis of vowel sounds. The insight resulting from those experiments suggests that:

- a more perceptually meaningful concept than the idea of formant is a new concept we call Perceptual Spectral Cluster, that identifies a localized concentration of spectral power containing important cues used in vowel recognition.
- PSCs can be suitably identified using smooth spectral envelope estimation based on short-pass lifting the real cepstrum.
- the two PSC-related features that were found to be more robust in vowel recognition are the center frequency of the first PSC, and ratio of the spectral power pertaining to the first two PSCs.

Our evaluation tests indicate that the new features have the potential to perform consistently better than the popular features based on the F1-F2 formant frequencies. Future research will involve a larger number of child and female speakers and will take into consideration the adaptive implication of perceptual aspects such as the influence of pitch and masking effects.

7. References