In *Voce, Cantato, Parlato. Studi in onore di Franco Ferrero*, E.Magno-Caldognetto, P.Cosi e A.Zamboni, Unipress Padova, pp. 143-148, 2003.

VOWELS: A REVISIT

Maria-Gabriella Di Benedetto

Università degli Studi di Roma La Sapienza
Facoltà di Ingegneria
Infocom Dept. Via Eudossiana, 18, 00184, Rome (Italy)
(39) 06 44585863, (39) 06 4873300 FAX, gaby@acts.ing.uniroma1.it

1. INTRODUCTION

Characterizing speech sounds in terms of acoustic parameters is a long-standing problem. As far as vowels are concerned, properties in the vowel acoustic waveform which are invariant with respect to speaker, language and phonetic context variations still remain to be identified.

When a vowel is produced the vocal tract can be modeled as a sequence of acoustic tubes resonating at particular frequencies, F1, F2, F3, called formants. The position of the tongue varies according to the vowel. As a consequence, the size of the acoustic tubes, the rigidity of the walls, and the tension of the vocal folds are modified and determine F1, F2, F3 values, as well as fundamental frequency F0. The acoustic model predicts the relative invariance of the formants of the extreme vowels [i, *a*,u] when, changing the speaker, the dimensions of the vocal tract are varied.

In previous research, vowels have been usually described by the first two formants, F1 and F2. As well known, F1 is related to height and F2 to backness, with reference to the position of the tongue during articulation. F1 vs. F2 patterns for Italian vowels were first published by Franco Ferrero [1]. Reference data for French can be found in [2], and for American English in [3,4].

Information related to formant time-variations is usually discarded since the F1 vs. F2 values are sampled within the steady-state. Formants, however, vary within the vowel, and a lack of evidence for a steady-state is often observed [5]. The problem is thus to understand the impact of F1 and F2 variations within the vowel on height and backness. This investigation was the focus of the present work. A subset of the entire set of American-English vowels was selected for the purpose of the study. This set was formed by the unrounded and non-diphthongized vowels of American-English. The analyzed vowels belonged to the Lexical Access database, developed in the Speech Group of the Massachusetts Institute of Technology, which contains 100 sentences uttered in a read-style mode. The same set of vowels, though in CVC syllables, had been already investigated several years earlier [5,6].

The paper is organized as follows. In section 2, a description of the lexical access database is given. Section 3 reports the measurement procedure and the results of the acoustic measurements. Results are discussed in section 4.

In *Voce, Cantato, Parlato. Studi in onore di Franco Ferrero*, E.Magno-Caldognetto, P.Cosi e A.Zamboni, Unipress Padova, pp. 143-148, 2003.

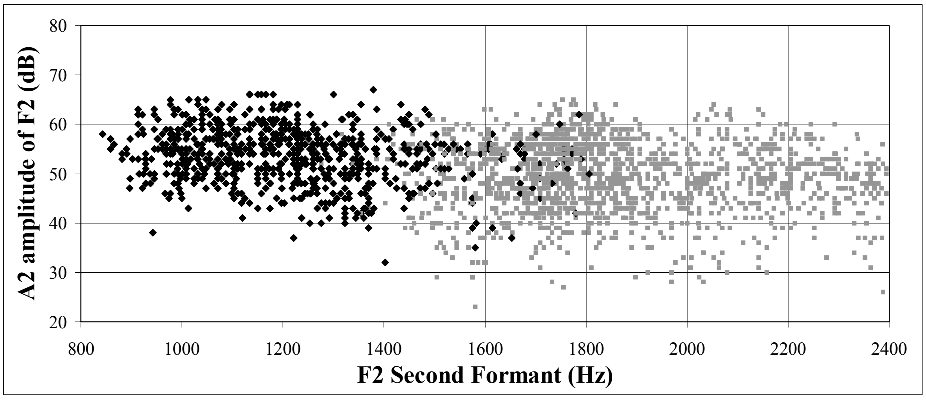

Figure 1 – F2 and A2 for all vowels and speakers. Front vowels in grey, back vowels in black

2. THE LEXICAL ACCESS DATA-BASE

The Lexical Access database was developed in the Speech Group of the Massachusetts Institute of Technology, Cambridge, USA. It consists of 100 sentences which were recorded in a soundproof room using high quality equipment. Four native speakers of American English, two males (k and m) and two females (s and j), uttered one repetition of each sentence. The speech materials were then converted in a numerical form (filtered at 7.5 kHz, sampled at 16 kHz, 12 bits/sample).

Five vowels [I,ε,æ,*a*,ʌ] were selected for this study. These vowels correspond to the set of monophthongal unrounded vowels of American-English. The selected vowels were either primary stressed or full vowels. Vowels occurring in nasal contexts were excluded.

3. ACOUSTIC MEASUREMENTS

Speech materials were analyzed using a software XKL [7]. This program computes DFT slices, a smoothed spectrum, and the LPC spectrum. The pre-emphasis filter coefficient was set at 0.99. Formants were obtained by using the smoothed spectrum with a 25.6 msecs window.
The following parameters were estimated: the first three formants (F1,F2,F3), their amplitudes (A1,A2,A3), the energy in the frame (A), and fundamental frequency (F0). These parameters were measured throughout the vowel, every 10 msecs.

In *Voce, Cantato, Parlato. Studi in onore di Franco Ferrero*, E.Magno-Caldognetto, P.Cosi e A.Zamboni, Unipress Padova, pp. 143-148, 2003.

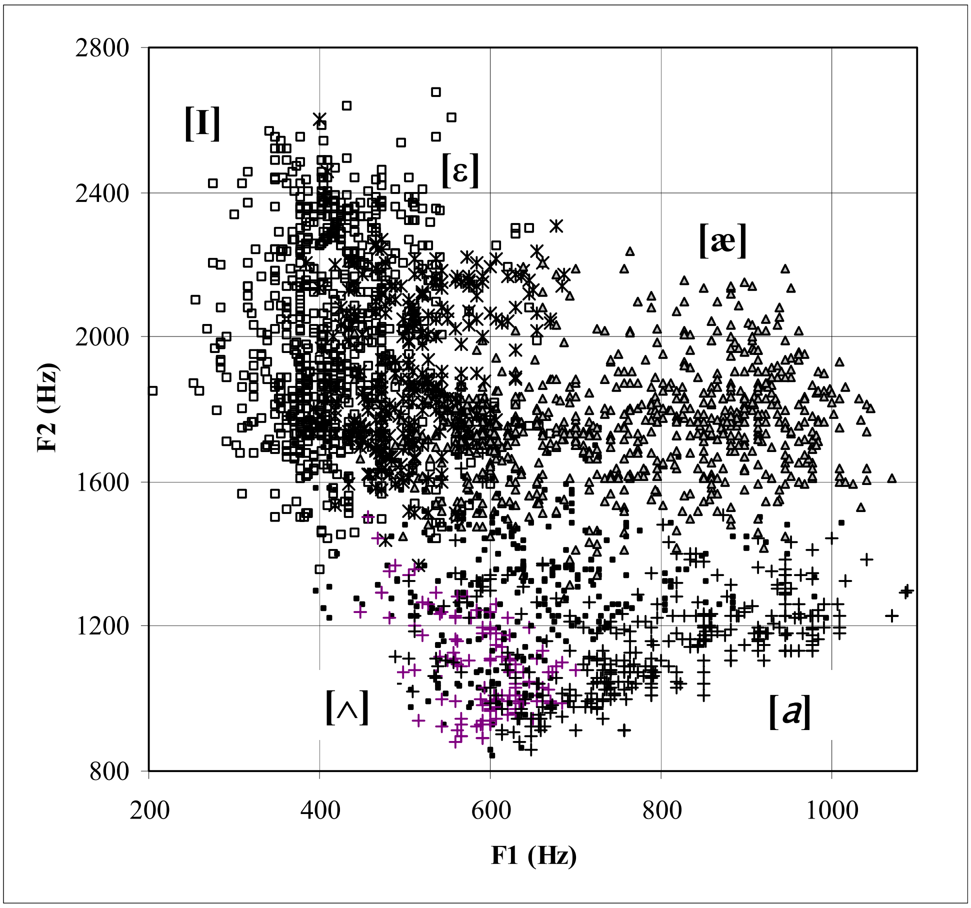

Figure 2 – Vowel representation for all speakers in F1vs.F2. Height is along the x-axis.

Results showing F2 and A2 values sampled throughout the vowel for all speakers and vowels are presented in Fig.1. Note that front vowels (grey dots) overlap with back vowels (black dots) in the 1400-1700 Hz region. Detailed analysis of the data showed, however, that there was no inter-speaker overlap. The overlap was mostly due to [ʌ] in function words, or in words such as "just" or "other" for which we can predict that contextual effects will tend to make the vowel front. A high F2 was also observed in a few tokens of the word "sudden" of speaker s. As a matter of fact, an F2 boundary set at about 1500 Hz may serve as an absolute boundary for separating back and front vowels of any speaker. Back vowels of male and female speakers had similar F2 values, and although front vowels had significantly higher F2 values for female speakers, the value of the F2 boundary is not affected; F2 normalization may not be necessary. This result confirmed similar findings in French vowels [2].

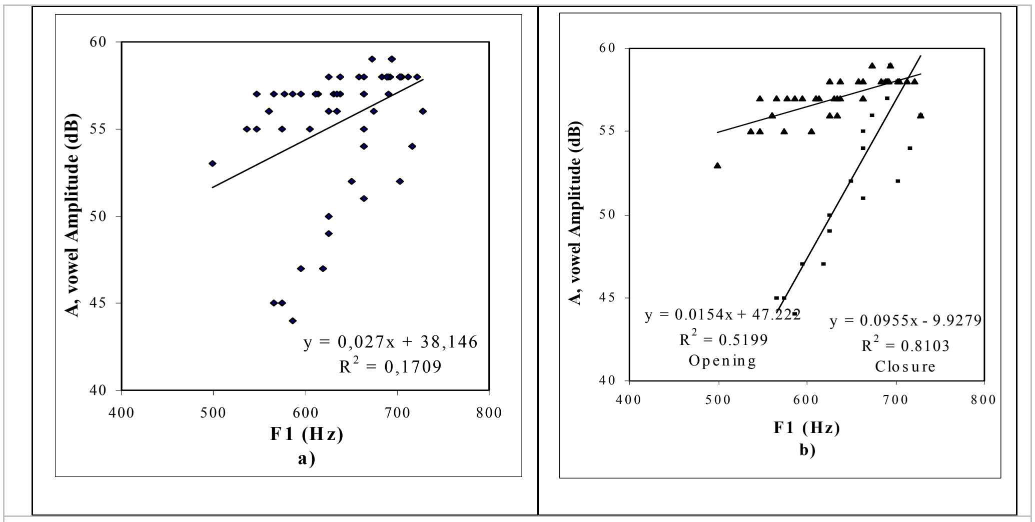

Figure 3 – Amplitude variation with F1 for a token of the vowel [*a*] (4 repetitions, speaker k). Figure 3a shows values and linear fitting of values in one cloud. Figure 3b shows the fitting when values are separated into opening and closing portions of the vowel.

We tested, however, the auditory parameter (F3-F2), in Barks, as suggested by Syrdal and Gopal for representing backness in American-English vowels [8]. Results on our data indicated that (F3-F2) did not perform better than F2 since more overlap was found with (F3-F2) than with F2. Therefore, F2 appeared as more robust than (F3-F2) with respect to variations of the formant pattern within the vowel.

Vowel areas in the F1 vs. F2 plane are shown in Fig.2. As regards height, note that vowels overlap significantly. In particular, the high vowel [I] overlaps with the non-high vowel [ɛ], the non-low vowel [ɛ] overlaps with the low vowel [æ], and the non-low vowel [ʌ] overlaps with the low vowel [*a*]. The overlap was also large for each speaker. F1 values for vowels with low F1 were similar for male and female speakers, while the opposite was true for low vowels. This observation confirmed the findings reported in [9], which analyzed the same vowels in CVC syllables. We tested the parameter (F1-F0), in Barks, which according to [8] reduces male-female differences (it has a normalization effect) and is more appropriate than F1 for representing height. Results confirmed previous investigations on the same vowels in CVC words [9] that the (F1-F0) distance actually increased male-female differences for high vowels since these vowels have similar F1 for male and female speakers. (F1-F0) reduced the male-female difference in low vowels for which female speakers have a significantly higher F1. Note however that this "compression" effect may not be necessary since low vowels of female speakers extended in a region which is not occupied by any other vowel. Therefore, similarly to backness, results indicated that F1 was more effective than using an auditory-based parameter such as (F1-F0). For back vowels, issues related to the interaction between F1 and F2 still need to be addressed (contrarily to front vowels which have F1 and F2 well apart).

Formant amplitudes A1, A2, A3 and the amplitude of the vowel A were then analyzed. The range of variation of A was about 20 dB. Results showed that A1, A2, A3 were all highly linearly correlated with A, and increased with A but with different rates. Overall, a spectral tilt was observed for some vowels but there was no systematic effect among speakers.

The analysis of F0 and formants in relation to amplitude A indicated that:

1. F0 was linearly correlated with amplitude A;
2. F1 was linearly correlated with amplitude A but the linear correlation coefficient was low;
3. F2 and F3 were not correlated with amplitude A.

These findings were in agreement with results reported for French vowels [2]. Note in particular that the rate of increase of F0 was here about 2.5 Hz/dB compared to 5 Hz/dB found for French vowels [2] which were however pronounced with different degrees of vocal effort. As regards F1, the rate of variation was here 5 Hz/dB compared to 3.5 Hz/dB of French vowels. These differences are small, also considering that different measurement tools were used.

The low correlation coefficient found for F1 was further investigated. Preliminary results indicate a possibility for a different rate in the opening portion (when F1 rises) compared to the closing portion (when F1 decreases). This result is illustrated in Fig.3 for a token of the vowel [*a*], speaker k; If all points of the trajectory are plotted in one cloud (Fig.3a), the correlation is low. Things however “straighten up” if dots are separated in two clouds (opening and closure, Fig.3b). Note the large increase in the correlation coefficient suggesting a different relation between F1 and A for the opening and closing gestures of the vowel.

4. CONCLUSIONS

Five vowels of American English [I,ε,æ,*a*,ʌ] belonging to sentences uttered in a read-style mode were analyzed. The vowels were represented by the first three formant frequencies (F1, F2, F3), their amplitudes (A1, A2, A3), the amplitude of the vowel (A), and fundamental frequency (F0), all sampled every 10 msecs, from the onset to the offset of the vowel.

The first question which was addressed was how to separate front and back vowels. Results indicated that an F2 boundary at about 1500 Hz was capable of separating well front and back vowels for both female and male speakers, and that the (F3-F2) distance in Barks did not achieve better separation. Moreover all F2 values within the F2 trajectory were on the right side of the boundary. Therefore, this parameter was robust with respect of time variations of F2. This finding also indicated that front-back classification might be performed very early in the vowel by the human processing system.

The second question which was addressed was how to classify vowels along height. When vowels were represented by F1, a large overlap between adjacent vowels was observed. This

In *Voce, Cantato, Parlato. Studi in onore di Franco Ferrero*, E.Magno-Caldognetto, P.Cosi e A.Zamboni, Unipress Padova, pp. 143-148, 2003.

overlap was due to both inter-speaker and intra-speaker variations. Using an auditory parameter such as (F1-F0) did reduce male-female differences for low vowels, but increased these differences for high vowels.

Finally, the relations between formants, formant amplitudes, and amplitude of the vowel were investigated. Vowel amplitude varied by an amount as large as 20 dB among the analyzed vowels. This fairly large range of variation may have an effect of formants themselves, and generally on the shape of the vowel spectrum. Results indicated that a spectral tilt was present in vowels with higher amplitude, i.e. there was a reinforcement of the high frequencies in the spectrum. Furthermore, F0 and F1 appeared to increase with amplitude, while F2 and F3 did not seem to be related to amplitude. As regards the relation between F1 and A, preliminary data suggested that the analysis should separate F1 onglide and offglide portions, and that the two portions might be characterized by different rates of variation.

Future research will be dedicated to a better understanding of joint variations of F1, F2, A1, and A2, and the possible interaction between F1 and F2 in back vowels in comparison to front vowels. As a general indication, we report that recent new findings on our data indicate that F1 might behave differently in back vowels than in front vowels as regards its relation with the relative amplitude of A1 to A2, i.e. the affiliation of F1 and F2 to front and back cavities. The explanation for this finding, whether it can be attributed to a production mechanism, remains to be clarified.

Acknowledgements

This work was partially supported by a grant of the Massachusetts Institute of Technology, Research Laboratory of Electronics. The author gratefully acknowledges Prof. K.Stevens for his support and encouragement.

REFERENCES

[1] Ferrero, F. “Diagrammi di esistenza delle vocali italiane”, Alta Frequenza, Vol 37, No 1, pp.54-58, 1968.

[2] Lienard, J.S. and Di Benedetto M.G. “Effect of vocal effort on spectral properties of vowels”, J. Acoust. Soc. Am., 106, 411-422, 1999.

[3] Peterson, G.E., and Barney, H.L. “Control methods used in the study of vowels”, J. Acoust. Soc. Am., 24, 175-184, 1952.

[4] Stevens, K.N., and House, A.S. “Perturbation of vowel articulation by consonantal context: An acoustical study “, J. Speech Hear. Res. **6**(2), 111-128, 1963.

[5] Di Benedetto, M.G. “Vowel representation: some observations on temporal and spectral properties of the first formant”, J. Acoust. Soc. Am., 86 (1), pp.55-66, July 1989.

[6] Di Benedetto, M.G. “Frequency and time variations of the first formant: properties relevant to the perception of vowel height”, J. Acoust. Soc. Am., 86 (1), pp.67-77, July 1989.

[7] Klatt, D.H. “M.I.T. SpeechVAX user’s guide”.

[8] Syrdal, A.K. and Gopal, H.S. (1986) “A perceptual model of vowel recognition based on the auditory representation of American English vowels”, J. Acoust. Soc. Am., 79, 1086-1100.

[9] Di Benedetto, M.G. (1994) “Acoustic and perceptual evidence of a complex relation between F1 and F0 in determining vowel height”, Journal of Phonetics 22, pp.205-224.