Vowels: a revisit

Maria-Gabriella Di Benedetto

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

ABSTRACT

Acoustic analysis of five vowels [I, ɛ, æ, *a*, ʌ] of the Lexical Access database, developped in the Speech Group of the Massachusetts Institute of Technology, was carried out. Vowels were represented throughout their duration by the first three formant frequencies, their amplitudes, the amplitude of the vowel itself, and the fundamental frequency, sampled every 10 msecs, from the onset to the offset of the vowel. Results indicated that an F2 boundary set at about 1500 Hz is capable of separating front and back vowels for both female and male speakers. The use of 'auditory' dimensions did not appear to achieve better separation. In the height dimension, the use of F1 produced a large overlap between adjacent vowels due to both inter-speaker and intra-speaker variations. Using auditory parameters such as F1-F0 did not lead to better high/non-high and low/non-low distinction, nor did it produce a male/female normalization effect. Analysis of the relations between formants and amplitudes revealed the presence of a spectral tilt in vowels with high amplitude.

Keywords: acoustic analysis, vowels

1. INTRODUCTION

This document reports the results of the analyses carried out at the Massachusetts Institute of Technology during Spring-Summer 1999. This work represents the outcome of an intense period of interaction with Kenneth Stevens, Stefanie Shattuck and the members of the Speech Communication Group at MIT.
The object of the work was to analyze a set of vowels in order to understand how to represent vowel features in terms of acoustic parameters. The analyzed vowels belonged to the Lexical Access database, developped in the Speech Group of the Massachusetts Institute of Technology, which contains 100 sentences uttered in a read-style mode by two female and two male speakers. 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-diphthonguized vowels of American-English. This same set of vowels, though in CVC syllables, had been already investigated ten years ealier [1,2,3].
The paper is organized as follows. In section 2, a description of the Lexical Access database is given. Section 3 reports the measurement procedure. Section 4 contains the results of the acoustic measurements. Results are discussed in section 5.

2. THE LEXICAL ACCESS DATA-BASE

The Lexical Access database, developped in the Speech Group of the Massachusetts Institute of Technology, consists of 100 sentences. The speech materials 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. During each recording session the recording level was kept constant. Therefore, energy of vowels of each speaker were comparable among each other. However, the recording level was not monitored from one recording session to the other; therefore, vowel tokens of different speakers could not be compared in terms of absolute energy and amplitude levels. The speech materials were then converted in a numerical form (filtered at 7.5 kHz, sampled at 16 kHz, 12 bits/sample) and stored on a computer.
Five vowels were selected for the study. These vowels were [I, ɛ, æ, *a*, ʌ] which correspond to the set of monophthongal unrounded vowels of American-English. This same set of vowels had formed the object of previous studies [1,2, 3]. As regards stress level, either primary stressed or full vowels were analyzed. Vowels occurring in a nasal context were excluded.

3. ACOUSTIC MEASUREMENTS

The vowel tokens were analyzed by means of the XKL program. This analysis tool is derived from the KLSPEC program developed by Dennis Klatt [4] under the VMS operating system. The KLSPEC software has been adapted to run on a PC under Unix. This software provides spectrum estimation at a given time, the spectrogram, and a variety of additional features. It allows visualization of the DFT slices of the signal, and computes a spectrogram-like spectrum (smoothed spectrum) and the LPC spectrum. The spectrogram-like spectrum is very useful for estimating formant frequencies in vowels. It allows, by means of an interpolation algorithm, to improve the accuracy over the resolution implied by the number of DFT samples over a given frequency range, which depends on the original filtering of the analog signal.
A measurement procedure, i.e. window length, preenphasis, and type of spectrum was selected. The preenphasis filter coefficient was set to 0.99. Formant measurements obtained by use of three

measurement methods were then compared. The three methods were LPC, smoothed-spectrum with window length equal to 6.4 msecs (short window), and smoothed-spectrum with window length equal to 25.6 msecs (long window). In the case of the smoothed-spectrum with short window, the spectrum was averaged over a period of time equal to 20 msecs, located around the time of interest.

As regards time sampling of the formants, results of previous studies [1,2,3] indicate that if a vowel must be represented by a single set of parameters (as for example formants) then the best sampling time is at F1 maximum. Another possible choice is sampling where the amplitude is maximum, which does not always correspond to the same time as F1 maximum (typically if the amplitude of second formant is higher then the amplitude of the first formant, as is often the case for back vowels). The F1 maximum sampling time was adopted, and F1 and F2 values obtained with the three measurement methods were compared. Results showed that there was a slightly larger spread in F2 values with the short window, and that the LPC F1 values were in general lower than the smoothed long window values, confirming previous studies [1,2,3]; The average difference was about 20 Hz (lower than the 40 Hz found in [1,2,3]). Note that the finding that LPC F1 values are lower that the smoothed spectrum F1 values is not in agreement with Hillenbrand and Neary [5].

4. ACOUSTIC MEASUREMENTS

The following parameters were estimated: the first three formants (F1, F2, F3), their amplitudes (A1, A2, A3), the energy in the frame (A), the fundamental frequency (F0). These parameters were estimated throughout the vowel, every 10 msecs. Therefore, each vowel was characterized by a matrix of values, where the rows were the different times and the columns were the parameter values.

Results of the acoustic analysis for the five vowels and the four speakers, in the backness dimension, are reported in Fig. 1, which shows F2 and A2 for all speakers and vowels on the same plot. Front vowels are in grey. Back vowels are in black. Vowel areas overlap in the 1400-1700 Hz region; Note however that it is not inter-speaker overlap and that F2 values are sampled throughout the vowel. Analysis of overlap details showed that the overlap was mostly due to [ʌ] in words such as “just” or “other”, i.e. words in which we can predict that contextual effects will make the vowel front, or in function words. A high F2 value was also observed for speaker s in some tokens of the word “sudden”.

Therefore, it appears that, in order to represent backness, normalization of the F2 dimension might not be needed. An F2 boundary at about 1500 Hz does not move across speakers and appears to be an absolute boundary, which separates back from front vowels.

The use of an auditory parameter such as (F3-F2) for representing backness in American-English vowels, as suggested by Syrdal and Gopal [6], was tested. Syrdal and Gopal [6] showed that for back vowels F3-F2>4 barks, while for front vowels F3-F2< 4 barks. Results obtained on our data using (F3-F2) indicated however that (F3-F2) did not represent backness better than simply F2. On the contrary, it seemed that more overlap was present in the (F3-F2) dimension than in the F2 dimension.

Thus fo F2 there is no real need for normalization since back vowels have similar F2 values for males and females, and although front vowels have significantly higher F2 values for female speakers, this does not affect the value of the F2 boundary. This result confirms similar findings on French vowels [7].

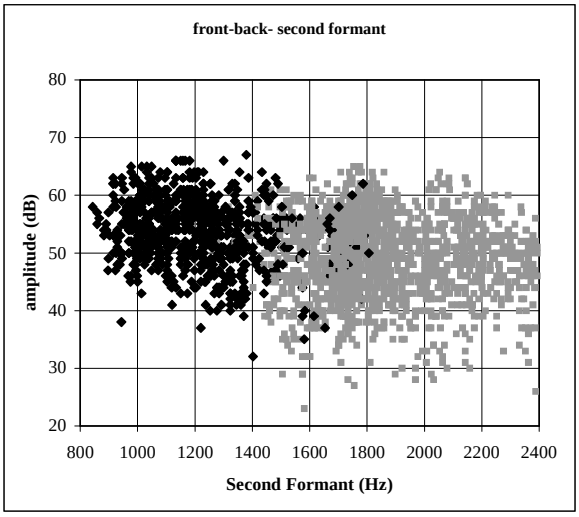


Figure 1 – F2 and A2 values for all vowels and speakers. Front vowels are grey, back vowels are black.

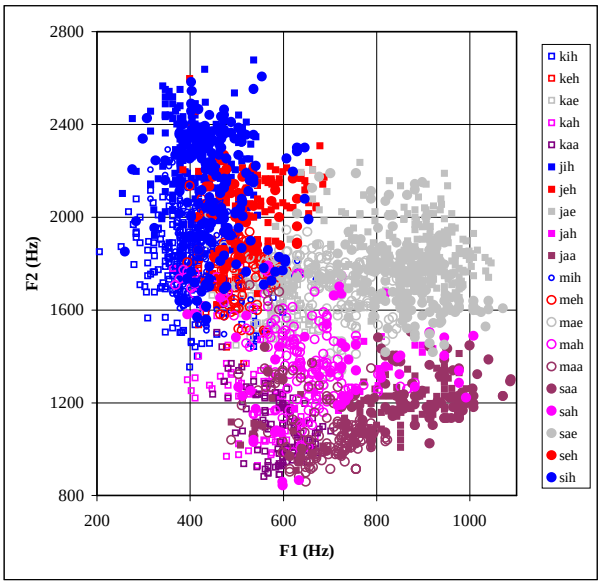


Figure 2 – Representation of the vowels of all speakers in the F1-F2 plane. The height dimension is along the x-axis.

Figure 2 shows vowel height along the x-axis. Note that vowels overlap significantly. In particular, the high vowel [I] overlaps with the non-high vowel [ɛ], the non-low [ɛ] overlaps with the low [æ], and the non-low [ʌ] overlaps with the low [*a*].

The overlap was large for each speaker individually. F1 values for vowels with low F1 were similar for male and female speakers, while the opposite was true for low vowels. This observation confirms the findings reported in [8], which analyzed the same vowels in CVC syllables.

As is the case for backness, an auditory parameter was proposed for representing vowel height [6], namely the distance between F1 and F0, F1 and F0 being in barks. According to [6], the F1-F0 parameter helps reducing the male-female differences (it has a normalization effect) and is more appropriate than F1 for representing height. Previous investigations on vowels in CVC syllables did not confirm this finding [8]. It was shown that the F1-F0 distance actually increased the male-female differences for high vowels because these vowels had similar F1 values for male and female speakers. For low

vowels, it reduced the male-female difference because these vowels had a significantly higher F1 for female speakers. However, it was noted that this "compression" effect was not needed since low vowels of female speakers extended in a region that was not occupied by any other vowel. These findings were also perceptually verified.
The F1-F0 parameter was tested on the vowels under analysis. Results confirm the findings of [8].
As regards height, the results of the present analysis indicate that, as for backness, the use of F1 alone is more effective than the use of the auditory-based parameter F1-F0. For back vowels, issues related to the interaction between F1 and F2 still need to be addressed (something which is not the case for front vowels which have F1 and F2 well apart).

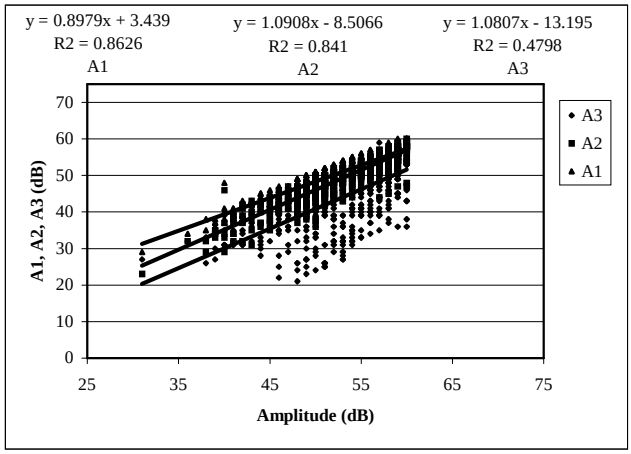
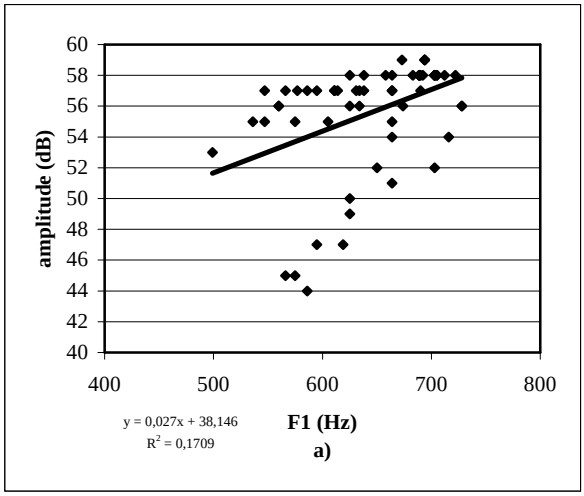


Figure 3 – A1, A2, and A3 variations with the amplitude of the vowel, for speaker k, when all vowels are considered in one set.

The amplitudes of the formants A1, A2, A3 were analyzed in their dependence upon the amplitude of the vowel itself, A. The vowels were analyzed throughout their duration, so that the results reported reflect also the behavior observed within a vowel and not only among tokens of the same vowel.
Results showed that the amplitudes of the first three formants were all highly linearly correlated to A. These amplitudes increased with A but in some cases with different rates. Overall, a clear spectral tilt was observed for some vowels but there was no systematic effect observed among speakers. For example, vowel [I] of speaker s had spectral tilt. Vowel [*a*] had a lower A3 amplitude than [∧]; this was observed for all speakers and should be investigated further.
The spectral tilt effect is shown in Fig.3, for speaker k. The range of variation of A was about 20 dB. In particular, note that A2 and A3 seem to increase with the same law, while A1 increases slower, and that. This observation is in agreement with the findings in [7] in which the range of amplitude variations was about 10 dB.

The analysis of the data concerning the formants and the fundamental frequency in relation to amplitude indicate that

1. F0 is linearly correlated with amplitude;
2. F1 is correlated with amplitude although the linear correlation coefficient is low;
3. F2 is not correlated with amplitude, although F2 should be plotted here in two sets, one for front and one for back vowels
4. F3 is not correlated with amplitude.



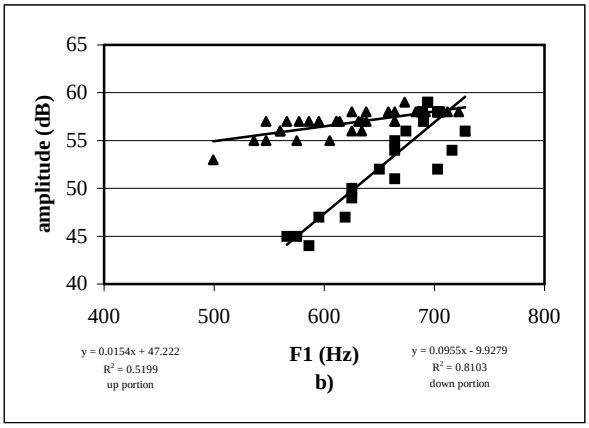


Figure 4 – Variation of amplitude with F1 for a token of the vowel [*a*] (pronounced in 4 repetitions by speaker k). Figure 4a shows the values and the linear fitting of the points when considered as only one cloud (note the low value for the linear correlation coefficient). Figure 4b shows the fitting when the points are separated into opening and closing portions of the vowel.

These findings are in agreement with the results reported on French vowels [7]. Note in particular that the rate of increase of F0 is about 2.5 Hz/dB compared to the 5 Hz/dB found on French vowels [7] which were however pronounced with different degrees of vocal effort. As regards F1, the rate of variation was 5 Hz/dB compared to 3.5 Hz/dB of the French vowels. These differences are small; Consider that different measurement tools were used.
The low linear correlation coefficient found for F1 was further investigated. Preliminary results indicate that there might be a different rate of increase-decrease in the opening portion (when F1 rises) compared to the closing portion (when F1 decreases). This result is illustrated in Fig.4 for a token of the vowel [*a*], speaker k. Note that in Fig.4a, if all points of the trajectory are plotted in one cloud, the correlation is low. Note, however, that things "straighten up" if the dots are separated into two clouds (up-trajectory and down-trajectory) as shown on Fig.4b. Note the large increase in the correlation coefficient value. This observation may explain why the F1 parameter had a low linear correlation with A; opening and closing portion should be considered separately, meaning that the relation between F1 and A may be different in the opening and closing gestures of the vowel.

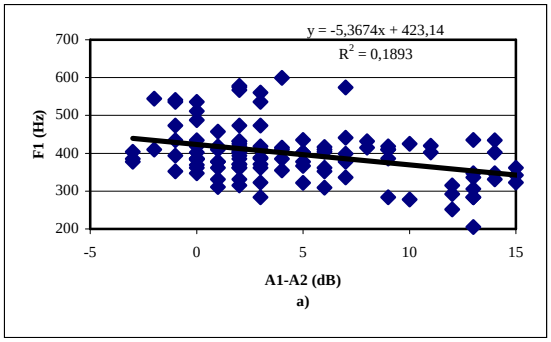
Joint variations of F1, F2, A1, and A2 were then investigated. The reason for investigating this relation lays in the possible interaction between F1 and F2 in back vowels in comparison to front vowels.

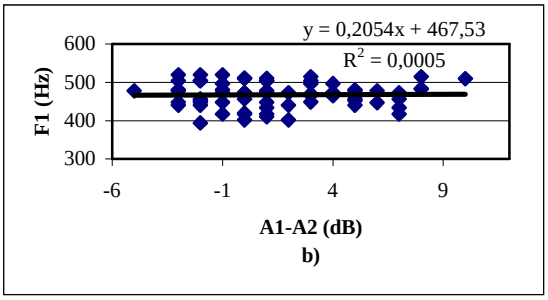
As previously mentioned, [*a*, ʌ] have very similar F1 values, and so far their main observed difference is the amplitude of F3 which is lower for [ʌ] than for [*a*]. As expected, F1 and F2 in back vowels were more difficult to measure because they were close in frequency. Primary stressed vowels, mainly for female speakers, were the most difficult ones, because the high amplitude made F1 higher and therefore closer to F2. In these cases, F1 and F2 often corresponded to two consecutive harmonics; the smoothed spectrum would in these cases locate one maximum in between the two harmonics.

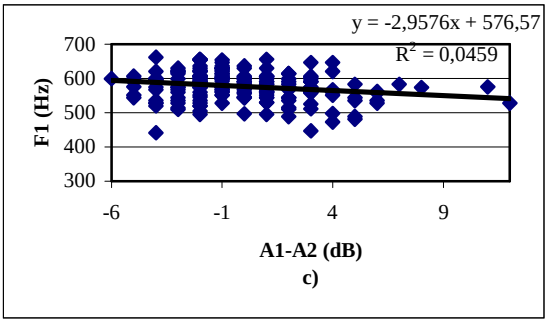
The question was to understand whether the high F1 values for [ʌ] were in some way connected to the amplitude of F2 relative to F1. Results are reported in Fig. 5 for speaker k, for vowels [I] (5a), [ɛ] (5b), [æ] (5c), [ʌ] (5d), and [*a*] (5e).

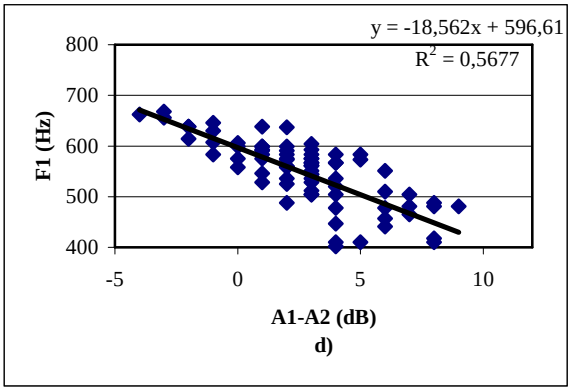
As can be noticed, F1 seems to be highly correlated with A1-A2 for back vowels, i.e. those vowels for which F1 and F2 are close in frequency. For front vowels, F1 does not show the same behavior. More specifically, F1 appears to be higher in back vowels when A2 is higher than A1, i.e. when more energy is placed on F2 than on F1. This effect can be observed for both [*a*, ʌ]. However the rate of variation seems different for these two vowels.

For the female speakers and speaker m, the same analysis is complicated by the fact that low vowels cross the 800-900 Hz region where an extra resonance due to the system below the glottis is present. If F1 traverses that region, the extra resonance superimposes to F1 and this results in a more complicated pattern to be understood. Analysis on the back vowel data of the female speakers indicate that if only those vowel which have an F1 lower than about 800 Hz were taken into account, then the same effect as the one observed above was present. This type of analysis might be helpful in better understanding how to represent height for [*a*, ʌ]. However, it is still unclear whether the observed effect is due to the measurement procedure, to production constraints, or to both.









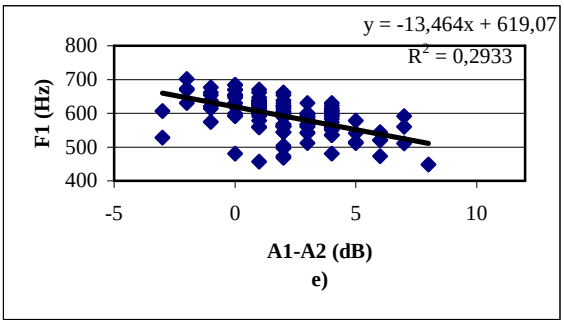


Figure 5 – F1 as a function of (A1-A2) for vowel [I] (5a), [ɛ] (5b), [æ] (5c), [ʌ] (5d), and [*a*] (5e).

5. CONCLUSIONS

The present work includes the acoustic analysis of five vowels in the Lexical Access database. The analyzed vowels occurred in different contexts, although vowels occurring in a nasal context were not part of the study. The vowels were represented throughout their duration by the first three formant frequencies (F1, F2, and F3), their amplitudes (A1, A2, and A3), the amplitude of the vowel itself (A), and the fundamental frequency (F0). The values of these parameters were sampled every 10 msecs, from the onset to the offset of the vowel.

The first question, which needed to be addressed, was how to separate front vowels from back vowels. To this regard, the results of the analyses indicated that an F2 boundary set at about 1500 Hz is capable of separating well front from back vowels for both female and male speakers. The use of the F3-F2 distance in barks did not appear to achieve better separation between these two sets of vowels. Furthermore, F2 by itself was capable of describing front and back vowels although its trajectory was not sampled. That is at any point within the F2 trajectory the F2 value was on the right side of the boundary. Therefore, this parameter is robust with respect of sampling time. This result also indicates that the front-back classification might be done very early in the vowel by the human processing system.

The second question, which needed to be addressed, was how to classify vowels along the height dimension. When vowels were represented by F1, a large overlap between adjacent vowels was observed. This overlap was due to both inter-speaker and intra-speaker variations. Using auditory parameter such as F1-F0 did not lead to better results in terms of high/non-high and low/non-low separation, nor did it produce a normalization effect between male and female speakers. In fact, it was noticed that while for low vowels the F1-F0 parameter did reduce the male-female differences, it increased these differences for high vowels. As a final remark, it should be noted that low vowels of female speakers extend in a region which is not occupied by any other vowel, and therefore there might be no need for shifting these vowel areas to lower frequency ranges.

Finally, the relations between formants, formant amplitudes, and amplitude of the vowel were investigated. Vowel amplitude did vary by an amount as large as 20 dB among the analyzed vowels. This fairly large range of variation might have an effect of formant values themselves, and generally on the shape of the vowel spectrum. Results indicate 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 suggest that the analysis should separate onglide portions from offglide portion of the F1, and that different rates might affect these two portions of F1 trajectory.

Finally, a first attempt was made to relate F1 values in back vowels to the relative amplitudes of F1 and F2. Results indicate that F1 might behave differently in back vowels and 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 remains to be clarified, whether it can be attributed to a production mechanism or to purely instrumental reasons related to the limitations of the frequency analysis tools

Acknowledgements

I wish to thank Ken Stevens for his invaluable comments as well as Stefanie Shattuck for discussions and suggestions, which greatly helped in the fulfillment of this work.
This work was supported by the Massachusetts Institute of Technology, Research Laboratory of Electronics formerly directed by Prof. J.Allen. The author gratefully acknowledges Prof. K.Stevens for the support received.
This work is dedicated to the memory of Franco Ferrero and his pioneering research on vowels.

References

[1] Di Benedetto, M.G. “An acoustic and perceptual study on vowel height”, Doctoral dissertation, University of Rome La Sapienza, 1987.
[2] Di Benedetto, M.G. “Vowel representation: some observations on temporal and spectral properties of the first formant”, Journal of the Acoustical Society of America, 86 (1), pp.55-66, July 1989.
[3] Di Benedetto, M.G. “Frequency and time variations of the first formant: properties relevant to the perception of vowel height”, Journal of the Acoustical Society of America, 86 (1), pp.67-77, July 1989.
[4] Klatt, D.H. “M.I.T. SpeechVAX user’s guide”.
[5] Hillenbrand, J.M. and Neary, T.M. “Identification of resynthesized /hVd/ utterances: Effects of formant contour”, Journal of the Acoustical Society of America, 105 (6), pp.3509-3523, June 1999.
[6] Syrdal, A.K. and Gopal, H.S. (1986) “A perceptual model of vowel recognition based on the auditory representation of American English vowels”, Journal of the Acoustical Society of America, 79, 1086-1100.
[7] Lienard, J.S. and Di Benedetto M.G. (1999) “Effect of vocal effort on spectral properties of vowels”, Journal of the Acoustical Society of America, 106, 411-422.
[8] 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.