Extraction of formants of oral vowels and critical analysis for speaker characterization

Odile Mella

To cite this version:

Odile Mella. Extraction of formants of oral vowels and critical analysis for speaker characterization. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Apr 1994, Martigny, Switzerland. pp.193-196. hal-01739608

HAL Id: hal-01739608
https://inria.hal.science/hal-01739608
Submitted on 21 Mar 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract

Methods for achieving automatic speaker recognition may be classified in two categories: pattern recognition based approaches and acoustic or phonetic feature extraction approaches. The former mainly apply and adapt to speaker recognition methods which have been already validated in automatic speech recognition such as template matching, hidden Markov models or neural networks. Such methods use implicitly interspeaker and intraspeaker variability of speech. Furthermore, in these approaches, the choice of the database on which speaker recognition methods are evaluated is not based on speaker recognition criteria. Most of them have been designed for speech recognition applications. So, their vocabulary is often restricted to common words like for instance the ten digits, which in English represent only half the English phonemes.

The second type of approaches examine linguistic units in order to extract features which are relevant for speaker characterization. Such approaches try to explicitly take into account the sources of interspeaker and intraspeaker differences. Most of the studies are related to the English language. We cannot here cite all of these experiments but they are described in [2]. Let us nevertheless quote U.G. Goldstein [?] and K.K. Paliwal [?] who investigate formants frequencies of respectively American English and British English vowels. In the French language, research in speaker characterization has mainly consisted of three studies. In 1971, P. Corsi carried out statistical analyses of several prosodic parameters and some segmental durations for 12 male speakers [?]. In [2], several phonetic, phonemic and prosodic features were analytically investigated from 5 male speakers’ utterances. By contrast, J.F. Bonastre in [2] did not try to directly extract relevant features. He showed the improvement of phonemes-based speaker identification when a similar context is used for reference and trial phonemes.

The purpose of our study is to examine the relative efficiency of the first three formants of the seven French vowels: /i/, /e/, /e/, /e/, /a/, /e/, /u/, with a preliminary neutral bilabial context /p/, /b/ and a subsequent lengthening context /R/.

O Mella is with the CRIN/CNRS & INRIA Lorraine, France. E-mail: mella@loria.fr.

Keywords — speaker characterization, speaker identification, vowel formants.

1. Introduction

2. Database

The seven French vowels are a part of a larger set of preselected acoustic and phonetic parameters which are assumed relevant for speaker characterization. To investigate these parameters 17 sentences are built and uttered four times by 18 male and 21 female speakers, coming from the same geographic region (Lorraine).

The database utterances are lowpass filtered to 6800 Hz and sampled at a rate of 16 kHz using a 12-bit analog-to-digital converter. We then hand-label 680 utterances of 10 male speakers. A broad phonetic transcription, including some infraphonemic labels such as burst and breathing, has been aligned by putting segment boundaries on the speech signal. To obtain a homogeneous labelling, to code as many as possible the speaker’s particularities and to allow for future feature analysis, we respect a set of strict rules for transcription and segmentation.

Among the 17 sentences, Table ?? only displays those including the triphones /p-vowel-R/ and /b-vowel-R/ but the original sentence numbering is kept.

3. Evaluation of reliable formant frequencies

The use of formants in speaker characterization require very reliable formant frequencies. Therefore, we compare the speakers in the same phonemic context but also in the same syntactic and semantic context. And above all, we use the knowledge of the vowel and of its context to establish a method to determine very reliable values of the three frequencies of the first formants.

Every “final formant” frequency is obtained from three "intermediate formant” frequencies which have been eval-
uated at three close locations in the vowel, at the vowel center and at 8 ms on either side of the center. The localization of the center depends on the vowel duration; at 80 ms from the beginning if the duration is greater than 160 ms, at the middle otherwise.

Every intermediate formant frequency is itself the result of an LPC-pole assignment method. Every formant (final or intermediate) is defined as a three fields structure: a frequency, $F_i$, a bandwidth $F_{i, bw}$ and a measure of disbelief $d_f$, which is itself composed of two fields, $d_{f_1}$ and $d_{f_2}$.

3.1 Determination of intermediate formants

The digitalized speech signal is preemphasized and, after applying a Hamming window, an autocorrelation analysis is carried out over frames of 256 samples (16 ms) to compute 18 LPC coefficients. Frequency-sorted poles with a bandwidth less than 1000 Hz are then extracted.

These poles are filtered a first time: when both poles have too close frequencies, only the pole with the lowest bandwidth is kept. For each vowel, in order to assign the remaining LPC poles to the three formant frequencies of the vowel, a second filtering classified them according to three definition domains $D(F_i)$ and removed those not falling into any domain. These vowel-specific domains are sequential intervals in which the first three formant frequencies of the vowel in the preliminary context /p/ and the subsequent context /R/ are assumed to be free for male speakers.

Three of the remaining poles are sequentially matched to the first three formants of the vowel according to an algorithm, which, for each formant $F_i$, takes into account first the number of poles included in $D(F_i)$ and then in priority order:
- the pole that has already been assigned to the formant $F_{i-1}$,
- the poles that could be assigned to the formant $F_{i+1}$,
- the relative candidates’ bandwidths,
- the frequency proximity between the poles candidates and a reference value for the frequency of $F_i$ for the given vowel.

The measure of disbelief $d_{f_1}$ is then set to 1 if the bandwidth of the retained pole is too large and to 0 otherwise. If no pole is available for a formant $F_i$, $F_i, fr$ is set to 0. Out of 5400 estimated formants, 1.3% had a null frequency, 95% stemmed from a single pole in $D(F_i)$, 3.5% from two poles and 0.2% from three poles.

3.2 Evaluation of reliable final formants

Table ?? shows how each final formant $F_i$ is computed from the three intermediate formants as a function of the proximity of their three frequencies. Two frequencies are said to be close if they differ at most of $E_i$. Three frequencies are said to be close if each pair of them differ at most of $E_i$. We have chosen for $E_i$ respectively 60 Hz for $F_1$ and 110 Hz for $F_2$ and $F_3$. We based our choice on the results about the accuracy of formant frequency measurements provided by R.B. Monsen and A.M. Engenbretson [?].

As can be seen in Table ??, the higher the value of $F_i, df_2$, the less reliable the final formant. Moreover, the final formant frequency is coded by a zero when it seems too unreliable, so that it can not be used to discriminate speakers. Out of 1800 formants only 26 are forced to 0.

3.3 Results about reliability of formants

For each studied triphonic formant reliability ratios (in %) are displayed in Table ???. First column mentions the vowel symbol followed by the sentence number. A very reliable formant has a null disbelief coefficient $d_{f_2}$ and a reliable formant a disbelief coefficient $d_{f_2}$ less than or equal to 1. An expected result that can be seen from the table is that the formant reliability depends on the syntactic and semantic localization of the vowel in the sentence. Given a vowel, the best reliability was reached for its stressed occurrences which appear at the end of words, syntagmes, and sentences. By contrast, its occurrences at the beginning of words or in grammatical words have less reliable formants.
for each vowel and for each type of combination of formant frequencies, a relevance indicator has been computed, the global speaker identification rate $R$.

4.2 Experimental results

4.2.1 Speaker identification modes

Generally, whatever the weighting, the vowel and the formant combination, taking into account the disbelief measure does not improve the speaker identification rate. Moreover, for every combination, this does not modify the order of relevance of the vowels. This is because the database sentences have been read and recorded in a quiet room and, as we showed previously, the formant measurements are almost all reliable. In what follows, only the simplest identification mode is assumed.

4.2.2 Vowel relevance for the formants $F_3$, $F_2$ and $F_1$

Figure 2 shows speaker identification rates for each formant and for each of the 15 vowels in its context. It can be seen from this table that, for speaker identification, the average performance of $F_3$ is better than the average performances of both $F_2$ and $F_1$.

![Figure 1: Speaker identification rate (in %) for each $F_i$.](image)

- **Formant $F_3$**
  
  With respect to $F_3$, the most effective vowel for speaker identification is /u/, except if it occurs at an unstressed localization in the sentence. More globally, we can notice that the occurrence $u_0$ in the preposition "pour" obtains the worst result whatever the formant combination.
  
  The /u/ vowel is followed by the two other rounded vowels /ɔ/ and /œ/ and by /a/. With respect to the rounding, /a/ is rather neutral while with the subsequent context /R/ is rather back like /ɔ/ and
It has been shown that $F_3$ is related to the labialization degree and that an increase in labialization causes a lowering of $F_3$ frequency. On the other hand, the uvular context /R/ leads to an increase in $F_3$ frequency. We conjecture that speakers could be discriminated by the way they round and in a less extent by their variability in coarticulation. Regarding earlier studies, $F_3$ of /u/ has been ranked as the second best feature for speaker identification but in the context /tu/ by M.R. Sambur [2] and as the sixth best feature by U.G. Goldstein [2]. On the contrary, in K.K. Paliwal’s study [2], the triphone /hud/ has obtained very poor performance. As far as we know, the relevance of $F_3$ for the /u/ vowel has not been individually experimented in French studies. As for $F_3$ of /a/, it has never been relevant in any of the English studies, at best ranked 10th out of 40 vowel formants in [2]. Then, G. Pérennou [2] has discriminated five male speakers by their way of using four allophones of the phoneme /a/ in a short text. But these allophones occurred in various consonant contexts while in our study the fixed back context should lead to a more homogeneous articulation among the speakers.

Figure ?? presents our more relevant vowels related to the non-uniform female/male formant frequency ratios ($k_j$) measured by G. Fant [2] and F. Lonchamp [2] and estimated by H. Traunmüller [2]. It can be seen that /u/ has the highest $k_3$ while /a/ has the lowest. The /a/ and /e/ have median $k_3$. However, the dispersion of $k_3$ among the vowels is less than that of $k_1$ and $k_2$.

- **Formant $F_2$.**

Regarding $F_2$, the most effective vowels are the high front /i/, the mid-high front /e/, the mid-low front /?,/ and the central vowel /o/. Over the ten male speakers, $F_2$ ranges for /i/ from 1850 to 2400 Hz and for /e/ to 1700 Hz. Three sources of interspeaker differences could account for the relevance of $F_2$ of front and close vowels. The first is the anatomical source: it has been showed that, in front vowels, $F_2$ is related to the back cavity [2]. The second comes from the way the speaker articulates; he indeed can increase $F_2$ without damaging the perception of adjacent vowels. The last arises from the discrepancies among the speakers in connection with the coarticulatory effect of the uvular context.

Considering earlier studies, $F_2$ of /i/ close to the French /e/ in the articulatory triangle, is immediately ranked after $F_2$ of /u/ in M.R. Sambur’s study. Further, among the vowels experimented in K.K. Paliwal’s study, $F_2$ of /i/, has showed the highest F-ratio. U.G. Goldstein, has not directly considered /i/, but she has ranked the maximum of $F_2$ during the dipthong /dɪ/ as the 6th most relevant feature, just before $F_3$ of /u/.

If we examine now Figure ??, we notice that our relevant vowels for $F_2$ correspond to the highest values of the $k_2$ ratio, especially for /i/ and /e/.

- **Formant $F_1$.**

Let us now turn to $F_1$. The speaker identification rates are globally small, but they split the studied vowels into two groups, the open vowels /a/ , /œ/ and /3/ which are relevant and the close ones /i/, /e/ and /u/ which are irrelevant, the vowel /ɛ/ being at the group boundary. The $F_1$ pertinence thus seems be related to the openness. It could have two explanations. Either the speakers could have different openesses, or the pertinence could come from an artefact of the LPC formant measurement which provide less accurate values for the low formant frequencies.

As for previous studies, U.G. Goldstein showed the relevance of $F_1$ for the dipthong /ɑr/ (ranked 2nd) and of the maximum of $F_1$ for the retroflex vowel /3/ (ranked 10th). It seems difficult to compare the results of both studies because of the shape of the tongue (retroflex tongue tip) during the utterances of /i/ and /3/. Nevertheless, it could be noted that /ɜ/ has $F_1$ and $F_2$ values close to the French /œ/ ones. Likewise, K.K. Paliwal showed that /ɜ/ and /3/ in a mid-low central vowel, are relevant for $F_1$ (respectively ranked 2nd and 3rd). With respect to the French language, it can be only again mentioned the results in [2] about the allophones of the /a/ vowel. Figure ?? shows that the /ɑr/, /a/ and /ɛ/ vowels have high $k_1$. By contrast, the irrelevant vowels in our study match those having lowest $k_1$.

4.2.3 Vowel relevance for the other formant combinations

Regarding the ($F_i,F_j$) combinations, ($F_2,F_3$) obtains the best speaker identification rates. In that case, the relevant vowels are /œ/ and /ɛ/ which decrease when the weighting gives less importance to $F_3$.

For ($F_1,F_3$) combination, the relevance of /œ/ vowel is effective whatever the weighting while that of /œ/ decrease when the weighting give less importance to $F_3$ (for instance, for /œ/ from 50% with no weighting to 44% for the weighting with the inverse of the reference formant value).

For ($F_1,F_2$) combination, it can be quoted the relevance of the /i/ vowel in addition of those of /œ/ and /ɛ/.

For ($F_1,F_2,F_3$), the speaker identification rate rises to 75% for /œ/ and for the euclidian distance. More globally, /œ/ and /ɛ/ keep their relevance for the euclidian distance. But, they decrease when the weighting gives less importance to high frequencies, to the advantage of /i/ and /3/.

When K.K. Paliwal used the four formants of the English vowels to discriminate 10 speakers, he showed that the more relevant vowels were /ɜ/ and /3/.

---

2The subsequent uvular context /R/ should lead to an articulation close to that of /ɑ/ but we have kept the transcription symbol /a/.
5. Conclusion

We have examined the relative efficiency of the first three formants of the seven French vowels: /ı/, /ɛ/, /ɛ/, /æ/, /a/, /ɔ/, /u/, with a preliminary neutral bilabial context /p/, /b/ and a subsequent lengthening context /R/. For that purpose, we have made sure that the formant frequencies were reliable. Thus, we have established a formant determination method based on the knowledge of the vowel and its context. With respect to the isolated formants, some of our results match those of earliest studies especially for \(F_2\) of the rounded vowels and for \(F_2\) of the high front vowels. Moreover, we have found a certain relationship between our relevant vowels and the non-uniform female/male formant frequency ratios \(k_i\). But our data and experiments are not enough to interpret this relationship, given the current anatomical interpretation of the non homogeneous \(k_i[?]\).