TPC-5 Signal Processing and System Control

Cepstral analysis for Formants Frequencies Determination Dedicated to Speaker Identification

Equips Treatment of the Signal and Medical Electronics, Laboratory of Electronics and the Information technology ' LETI ', ENIS, Sfax's university – Tunisia

E-mail: tgdorra@yahoo.fr, mondher_frikha@yahoo.com

Adress:

Mondher FRIKHA
Rte Menzel Chaker Km 1.8
Rue des Roses
Sfax 3072
Tunisia
Phone: 216-74 859 660
Abstract

In this paper, we present a technique of parameterization of the speech based on the cepstral analysis, for the extraction of the first four formants F1, F2, F3 and F4 with the aim of a biomedical application. Indeed, such analysis, supposed linear, assures the speech signal deconvolution. It allows to separate the contribution of the vocal tract, i.e. the formants frequencies, and the one of the vocal cords responsible for the fundamental frequency.

The technique applied to some vowels extracted from the TIMIT database, allows to identify the variations interlocutors of the formants frequencies according to the sex and of the region. Variability interlocutor is a major phenomenon in speech recognition because a speaker remains recognizable by the timbre of his voice in spite of a variation which can sometimes be significant. Results so obtained allow to notice the variability of the formants frequencies of a vowel pronounced by various speakers. So, several scenarios were tested to know:

- A vowel pronounced by four men and four women who lived in the same region
- A vowel pronounced by four women of the same region
- A vowel pronounced by eight men who lived in different regions.

I. Introduction

The use of formant frequencies in speech analysis and modelling is appealing in principle due to their close relation to the vocal tract geometry. Unfortunately, reliable formant frequency estimates are very difficult to extract from the speech wave. However, several studies have shown the existence of an approximately linear relationships between formant frequencies and other spectral representations as the formants extraction by cepstral analysis, linear prediction coefficients or by linear prediction based cepstral coefficients ... [1],[7],[8],[9]. For accurate estimation, it is therefore necessary to perform a deconvolution to separate the contribution of the vocal tract from that of the source. Cepstral analysis and linear prediction are the two widely used techniques [6].

In this sense, the technique presented in this paper is based on the application of the cepstral analysis in order to extract the frequencies of the first four formants of some vowels, obtained from the TIMIT database, pronounced by speakers of different sexes and from varied regions. The comparison of the obtained results, allows to show the differences and the interlocutors variations of the formants frequencies according to the sex and the region.

The remainder of this paper is organised as fallows: The next section describes the basis notions of the cepstral analysis theory. The following section presents the results of the application of this technique for the formant extraction of some vowels pronounced by different speakers. We explain indeed these results by showing the interlocutor variability. A conclusion is provided in the last section of this paper.

II. Cepstral analysis

Cepstral analysis is at present very used in speech recognition since its introduction in the early 1970’s [3]. It is an analytical method of the vocal signal based on a modeling. Indeed, the major defect of the FFT for the calculation of the spectrum is that it does not take into account the inter-modulation source/conduit which makes difficult the measures of formant Fi and the fundamental F0 (indicated often by ‘ pitch’) characteristics respective of the vocal tract and the source.

Cepstral Smoothing or cepstre is a method aiming at the separation of their respective contributions by deconvolution. For it, the vocal signal Sn is supposed produced by an excitation signal gn (glottal source) crossing a linear passive system with impulse response hn (vocal tract) [ 1 ].

The vocal signal Sn results from the convolution of the source (excitation) and the vocal tract (modeled by a passive filter) such as:

\[ Sn = gn \ast hn \]
We make the hypothesis that the excitation source $g(n)$ is either a sequence of impulses (periodicals, of period $T_0$, for voiced sounds), or a white noise, according to the model of production.

A Z-transformation allows to transform the convolution into product:

$$S(z) = G(z) \cdot H(z)$$

If we take the logarithm of the module of the Z-transformed signal, the product will be converted into a sum and we have:

$$\log |S(z)| = \log |G(z)| + \log |H(z)|$$

Next, by inverse Fourier transform of the log spectrum, the cepstrum is computed [3]. In the practice, the Z transform is replaced by a FFT which possesses the same properties of linearity that Z. The expression of the cepstrum is so:

$$\zeta(n) = \text{FFT}^{-1}(\log(\text{FFT}(s(n))))$$

It displays the ripples and “waveform” of spectral representation in terms of “quefrencies”, the unit of which is a second.

At this state, the excitation and the vocal tract shape are superimposed, and can be separated using conventional signal processing such as a temporal filtering (liftering) [3]. In fact, the low order terms of the cepstrum contain the information relative to the vocal tract. This contribution becomes unimportant from a sample 0 (0 corresponds to the fundamental frequency $F_0$ [1]). The visible periodic peaks beyond 0, reflect the impulses of the source.

The last stage is the cepstral shape smoothing which aims to reconstitute the signal of the vocal tract and to smooth the spectral shape to extract the formants. The smoothed specter of amplitude can be obtained by the following scheme [1, 4]:

**Formant Detection:**

The vocal tract presents some appropriate pulsations very apparent in the specter of the acoustic signal. These appropriate frequencies constitute the formants of the vocal signal.

For a vocal tract of which length is of the order of 17 cms, we can observe 3 or 4 formants between 100 and 5000 Hz [2]. The first three formants are indispensable to characterize the vocal specter; the two following ones are useful for a qualitative synthesis.

Indeed, the smoothed specter and embarrassed of the contribution of the source contains only information on the vocal tract and in particular its peaks correspond to the resonance of the vocal conduit: formants. Then, frequencies measured in the maximums of envelopes are the frequencies of formants.

However, it is only about a valuation of the frequencies of formants because the specter is an estimated specter.

**III Results**

We applied the cepstral method described in previous to represent the shape of the speech specter and to extract the formants. The used speech signals were arose from the corpus TIMIT where we chose to take various vowels of a reference sentence in American English, common to several speakers of the two sexes and of different regions.

To differentiate the formants of vowels pronounced by the same sex, four women of similar ages, sizes and cultures and belonging to the same region was tested. On the other hand, to differentiate the formants of different sexes, we used four men and four women who lived in the same region. Finally, to differentiate the formants of two different regions, a comparison has been done between four men of one of the district in relation to four men of the other region. These men have similar ages. In every test, we chose one phoneme common to all speakers. The chosen regions are of United States. The first region is New England noted R1 and the second is Northern noted R2. We call region of a speaker the geographical area of the U.S. where he lived during his childhood years[10].

The obtained results are summed up in the below tables.
Experience 1

The table 1 presents a comparison of the frequencies of the first four formants of the vowel "ih" (of the sentence sa1du corpus TIMIT) pronounced by four feminine subjects belonging to the region R1.

<table>
<thead>
<tr>
<th></th>
<th>ih</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Woman1</td>
<td>479.16</td>
<td>1937.533</td>
<td>2802.133</td>
<td>3687.533</td>
<td></td>
</tr>
<tr>
<td>Woman2</td>
<td>536.458</td>
<td>2182.316</td>
<td>2541.683</td>
<td>3661.466</td>
<td></td>
</tr>
<tr>
<td>Woman3</td>
<td>500</td>
<td>2250.05</td>
<td>2937.5</td>
<td>3000</td>
<td></td>
</tr>
<tr>
<td>Woman4</td>
<td>401.041</td>
<td>2270.866</td>
<td>2916.833</td>
<td>3614.6</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Formants of the vowel 'ih' pronounced by four feminine speakers (R1) (sentence extracted from the corpus TIMIT)

To show more clearly the differences between the formant frequencies, we represent the following histogram:

Fig 1: Formants of the vowel 'ih' pronounced by four women (R1)

A variation in the frequency of every formant of the same phoneme is noticed for the different women. The weaker variation is observed for F1 and will be more important for the other formants by order. Then plus the formant has a superior order plus the difference of this formant frequency becomes higher. These variations can be explained by the fact that the identity of the speaker is transported by the anatomy of the vocal length of the person, tone, and intonations (indicated by the dynamics of formants).

Indeed, the main cause of differences interlocutors is of physiological nature. Speech is mainly produced by the vocal cords, which generate a sound in a basic frequency, the fundamental. This basic frequency will be different from an individual to another one and more generally from a sex to another. This sound is then transformed through the vocal conduit, bounded in its extremities by the larynx and the lips. This transformation, by convolution, allows to generate different sounds. The vocal conduit is of shape and length variable according to the individuals. The possible convolutions will be so different. Then, the same phoneme will be able to have very different acoustic realizations and different formant frequencies from an individual to another.

Experience 2

To compare the frequencies of the different formants of the same phoneme pronounced by speakers of different sexes belonging to the same region, we give the table 2 which presents the first four formants frequencies of the vowel 'aa' of the sentence 'SA1' extracted from the database TIMIT, pronounced by subjects of different sexes and belonging to the region R2.

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man 1</td>
<td>718.75</td>
<td>1510.41</td>
<td>2156.266</td>
<td>3677.1</td>
<td></td>
</tr>
<tr>
<td>Man 2</td>
<td>781.28</td>
<td>1431.28</td>
<td>2000</td>
<td>3162.52</td>
<td></td>
</tr>
<tr>
<td>Man 3</td>
<td>648.437</td>
<td>1351.6</td>
<td>2281.3</td>
<td>3132.85</td>
<td></td>
</tr>
<tr>
<td>Man 4</td>
<td>708.333</td>
<td>1284.733</td>
<td>2041.666</td>
<td>3538.211</td>
<td></td>
</tr>
<tr>
<td>Woman1</td>
<td>776.041</td>
<td>1937.516</td>
<td>2119.8</td>
<td>3760.433</td>
<td></td>
</tr>
<tr>
<td>Woman2</td>
<td>786.458</td>
<td>1750.016</td>
<td>2552.1</td>
<td>3744.816</td>
<td></td>
</tr>
<tr>
<td>Woman3</td>
<td>718.75</td>
<td>1731.28</td>
<td>2187.54</td>
<td>3906.26</td>
<td></td>
</tr>
<tr>
<td>Woman4</td>
<td>680</td>
<td>2210.95</td>
<td>2320.325</td>
<td>3851.6</td>
<td></td>
</tr>
<tr>
<td>Average (man)</td>
<td>714.2</td>
<td>1394.505</td>
<td>2119.808</td>
<td>3377.670</td>
<td></td>
</tr>
<tr>
<td>Average (woman)</td>
<td>740.3124</td>
<td>1907.4406</td>
<td>2294.941</td>
<td>3815.777</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Formants of the vowel 'aa' pronounced by speakers of different sexes from the region 2 (sentence extracted from the corpus TIMIT)

To better show the formant frequencies differences between various sexes, we calculated the frequency average of every formant for every speaker’s sex (table 2) and we drew the corresponding histogram (figure 2). So, we calculated the coefficient ki, which is equal to the report of the frequency of the formant Fi of a feminine speaker by the frequency of the formant Fi of a male speaker.
We notice that the coefficient $k$ is greater than 1 and afterward, the formants frequencies of a feminine subject are more raised than those of a male subject.

Indeed, male voice is more grave than feminine voice, frequency of fundamental being weaker. The vocal conduit is of shape and length variable according to the individuals and, more generally, according to the kind and the age. So, the adult feminine vocal conduit is, on average, lower 15% in length than that of an adult male vocal conduit [1]. The vocal conduit of a very young child is naturally lower in length than that of the adult [5].

For a constant shape, the feminine formants are generally 15% more raised than those male [1]. So, the larynx at the man is placed lower than that at the woman and afterward the pharynx of the man is proportionally longer. This in for effect of resources of different discards in the formants frequencies of the two sexes according to formants and vowels because of their degree of membership with the pharyngeal part of the conduit.

**Experience 3**

To show the influence of the region on the formants frequencies of the speakers, we present table 3 that gives a comparision of the frequencies of the first four formants of the vowel "i'y" (of the sentence SA1 of the corpus TIMIT) pronounced by eight male subjects. Every four speakers belong to a region (R1 or R2). So, for every region, we calculated the frequencies average of every formant for the four speakers (table 3) and we drew the corresponding histogram (figure 3).

**Table 3: Formants of the phoneme 'i'y' pronounced by male speakers from two different regions (sentence extracted from the corpus TIMIT)**

<table>
<thead>
<tr>
<th>Region</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man 1</td>
<td>385.41</td>
<td>2291</td>
<td>2541</td>
<td>3500</td>
</tr>
<tr>
<td>Man 2</td>
<td>343.75</td>
<td>2385</td>
<td>2800</td>
<td>3729</td>
</tr>
<tr>
<td>Man 3</td>
<td>357.14</td>
<td>2325</td>
<td>2772</td>
<td>3334</td>
</tr>
<tr>
<td>Man 4</td>
<td>333.33</td>
<td>2114</td>
<td>2604</td>
<td>3229</td>
</tr>
<tr>
<td>Man 1</td>
<td>348.958</td>
<td>2385.433</td>
<td>2385.433</td>
<td>3869.816</td>
</tr>
<tr>
<td>Man 2</td>
<td>359.357</td>
<td>2078.15</td>
<td>2609.4</td>
<td>3468.75</td>
</tr>
<tr>
<td>Man 3</td>
<td>302.083</td>
<td>2218.76</td>
<td>2385.433</td>
<td>3197.933</td>
</tr>
<tr>
<td>Man 4</td>
<td>343.75</td>
<td>2093.766</td>
<td>2572.933</td>
<td>3625.033</td>
</tr>
<tr>
<td>Average (R1)</td>
<td>354.908</td>
<td>2278.75</td>
<td>2679.25</td>
<td>3448</td>
</tr>
<tr>
<td>Average (R2)</td>
<td>338.537</td>
<td>2194.027</td>
<td>2488.299</td>
<td>3540.383</td>
</tr>
</tbody>
</table>

Fig 3: Average of formants frequencies of the phoneme 'i'y' pronounced by male speakers from two different regions

Differences in the frequencies of the formants of the same vowel pronounced by speakers of the same sex and different regions are noticed. We remark again that the difference of the first formant frequency is the lowest. These differences will be clearer by pronouncing a word or a complete sentence. This is, to better show the effect of the accent, which transports changes of pronunciation in the shape of replacement, abolition or insertion of unities of the phoneme in the "standard" transcription of words and through systematic changes in the accentuation of the tone, lasted, intonation, and stress.
Indeed, variability interlocutor finds its origin in the differences of pronunciation which exist within the same language and which establish the regional accents. These differences will be careful all the more easily as a community of language will occupy a very vast geographic area, without even taking into account the possible international brilliance of this community and so of the probability that has the language to be used as second or, worst, third language by an individual of foreign mother tongue. They’re also, phonetic definition just as much as a strict definition of a vocabulary or of a grammar can be put in badly.

Finally, the analysis of the results of these three realized experiences reveals at once variability in the frequencies of the formants of the same vowel pronounced by speakers of various sexes and of varied regions. These results are justified by the presence of three sources of variability such as physiological differences between the speakers, the effects of the coarticulation and the variable latitudes of realization in the linguistic plan [1].

IV Conclusion

We presented in this work an application of a technique of parameterization of the speech based on cepstral analysis for the extraction of the formants of vowels. Several speakers of different sexes and regions realized the application. Indeed, this technique allows putting in evidence the differences of the formants frequencies of a vowel according to the sex and the region. For it, we presented the results of tests elaborated at the beginning by speakers of the same sex, then by speakers of different sexes and belonging to the same region and finally, by speakers of the same sex but that different by the region. Indeed, we explained these formants frequencies variations. This study allows including all the differences of low level, which can exist in a message nevertheless bearer of the same information. However, these differences in the signal explain all the difficulty which can be engendered, and all the errors which can be provoked, by methods operating only a general comparison enters a speech signal to interpret and its exact phonetic definition, that this definition is registered under shape of rules or under shape of a corpus of reference forms [5].

V. References

[3]: Picone J. , «signal modeling techniques in speech recognition», proceedings of the IEEE, 81(9):1215-1247, 1993