APPLICATION OF THE EMD DECOMPOSITION TO DISCRIMINATE NASALIZED VS. VOWELS PHONES IN FRENCH

M.M. SAIDI$^{1,2}$
$^1$IMS Research Group
SUPELEC-Metz Campus
2 rue Edouard Belin
57070 Metz, FRANCE
email: MohamedModhaffer.Saidi@supelec.fr

O. Pietquin$^1$ and R. André-Obrecht$^2$
$^2$SAMova Research Group
Paul Sabatier University
IRIT UMR 5505
31062 Toulouse, FRANCE
email: Olivier.Pietquin@supelec.fr and obrecht@irit.fr

ABSTRACT
This work deals with the application of the Empirical Mode Decomposition (EMD) with the goal of showing it is capabilities and limitations when applied to nasalized or oral vowel phones. The method used in this study consists in three classical stages: signal preprocessing, feature extraction and decision. Firstly, the speech signal is decomposed using the EMD method to extract the three first Intrinsic Mode Functions (IMF). Then, Mel-Frequency Cepstral Coefficients (MFCC) are extracted from these IMFs or a (partial) sum of it. Finally, Artificial Neural Network (ANN) is used to distinguish nasal vowels from oral vowels in French (French database Bref80). Besides the fact that this study resulted in a significant improvement in the level of discrimination, when we use our method compared to the standard application of MFCC to the original signal. It has also allowed us to know which IMFs allows to better characterize the nasal vowels from the oral vowels.

KEY WORDS
EMD, IMF, MFCC, ANN, Nasalized phones, Vowels.

1 Introduction

In the last decade, there is several studies dealing with the extraction of articulatory features from the speech signal in view to their integration in recognition or synthesis speech systems [1], [2]. These features are usually points or places of articulatory movements. From this characteristics one can distinguishes the nasal or the nasalized phenomena. Nasal consonants are produced when the nasal and oropharyngeal resonators couple. Nasalized phones occur when the oral cavity is not completely closed by tongue or lips, air flows through both nasal and oral cavities. In the general case, production of nasal or nasalized phones seems to be equivalent from an articulatory point of view. This is due to the fact that almost nasals was voiced. The coupling between the oral and nasal cavities implies that the acoustic spectrum of nasal vowels (also called nasalized phones) is more complex than the spectrum of its oral counterpart. In fact, these two cavities affect each other and when they resonate together it results into a loss of amplitude or antiresonance at certain frequencies. This implies a creation of additional poles and zeros. This makes the study of the nasality very complex because the resulting signal cannot be considered as autoregressive like assumed by most feature extraction methods.

Due to this, several studies have been made in view of extracting or creating new parameterizations that can characterize the nasality of a speech signal [3]. The main drawback of these studies is the fact that they focus on the features extraction and not on the signal specificity.

In previous work [3], the capability of a new decomposition method called Empirical Mode Decomposition (EMD) was studied, with a standard LPCC parametrization to characterize a nasal character present in speech signal. In this work, our purpose is to apply this decomposition (i.e. EMD) to the nasalized phones. This technique will allow us to decompose the signal in its natural bases. Then we will apply a standard parametrization MFCC to some output with the goal of distinguishing nasalized vs. all vowels in the French language. If this system performs well one can envision to create a system of nasal/vowel recognition. This type of system is very helpful for the Automatic Speech Recognition (ASR) system.

The remainder of this paper is organized as follows. The next section introduces the EMD method with some qualitative results illustrating the interest of this approach. In section 3, the proposed method to decide between nasal vowels or oral vowels is presented; some results are discussed, before concluding and proposing some perspectives.

2 Empirical mode decomposition

The aim of the empirical mode decomposition is to study the local oscillation of the input signal [4]. For this purpose, this technique consists in decomposing the signal sequentially into Intrinsic Mode Functions (IMF): $d_i : \mathbb{R} \to \mathbb{R}$. Each $d_i$ is an oscillatory function with locally zero-mean everywhere; it is modulated in magnitude and frequency. The IMF have necessarily one zero-crossing between two consecutive extrema.

Practically, if we consider two consecutive local minima of the signal $S$ (at time $t_1$ and at time $t_2$), the first IMF located between these two times $\{d_i[S](t), t_1 \leq t \leq t_2\}$
appears as a local high frequency contribution (i.e. local detail). We note the complementary contribution \( r_1[S](t) \) and call it local trend (it can be seen as a residue of the first stage of the decomposition). So, the signal decomposition can be written as follows:

\[
S(t) = r_1[S](t) + d_1[S](t)
\]  

(1)

To process the whole signal, the procedure will be reproduced on the residue \( r_1 \). It is repeated until the last residue \( r_k \) has no extrema anymore or until some stopping criterion is met. Thus, we can write this decomposition for any depth \( K \), as follows:

\[
S(t) = r_K[S](t) + \sum_{k=1}^{K} d_k[S](t)
\]  

(2)

\( K \) is the number of IMFs obtained from the decomposition of \( S \) by EMD and \( r_k \) is the \( K^{th} \) residue obtained by analyzing the \((K-1)^{th}\) residue, associated to the \( K^{th} \) IMF \( d_K \).

The corresponding algorithm is as follows:

1. Initialization: \( r_0(t) = S(t) \), for \( k = 1 \)

2. Extract the \( k^{th} \) IMF, noted \( d_k(t) \) (step called sifting process [4])
   
   (a) Initialization: \( h_0 = r_{k-1}, j = 1 \)

   (b) Extract local minima and maxima of \( h_{j-1}(t) \)

   (c) Interpolate between minima (resp. maxima), to obtain the “envelope” \( EnvMin_{1-j}(t) \) (resp. \( EnvMax_{1-j}(t) \))

   (d) Compute the mean “envelope”:

   \[
m_{j-1}(t) = \frac{1}{2}[EnvMin_{1-j}(t)+EnvMax_{1-j}(t)]
\]

   (e) \( h_j(t) = h_{j-1}(t) - m_{j-1}(t) \)

   (f)  
   - if stopping criterion is met then \( d_k(t) = h_j(t) \)
   - else goto (b) with \( j = j + 1 \)

3. \( r_k(t) = r_{k-1}(t) - d_k(t) \)

4.  
   - if \( r_k(t) \) has at least two extrema, goto (2) with \( k = k + 1 \)
   - else the decomposition ends with \( r_k(t) \) being the residue of the decomposition

One can notice that the principle of the EMD decomposition differs from the original algorithm [4] by the addition of the sifting process. Indeed, this step allows to have symmetric envelope (i.e. EnvMin and EnvMax). The stopping criterion used in 2.f can be the standard deviation done in \( h_j \) or a maximum predefined number of iterations.

An example of the EMD decomposition applied to a sum of cosines is shown in figure 1. The signal is composed of three pure frequencies (i.e. 50Hz, 500Hz and 1000Hz); we can see the importance of the sifting phase.

Figure 2 illustrates the spectral effect of the EMD decomposition on the same sum of cosines; the first three IMFs correspond to each fundamental frequency. In other words, each Intrinsic Mode Function represents a frequency band.

3 Experiences

To quantify the interest of the EMD to characterize the nasal feature in the context of vowels, we have defined a classification system and we use the BREF80 [5] database. The choice of this basis is motivated by the existence of 3 different nasal vowels in the French language (i.e. [œ], [ø] and [œ]), noted N (according to the authors of [6], the nasal vowel [œ] begins to disappear). We use all oral vowel phones in the database (i.e. [a], [i], [y], [œ], [o], [œ], [e], [o] and [u]), noted V. This speech database is multi-speaker (male and female), the sample rate is 16kHz. A phonetic label is available in [7].

3.1 Method

To deal with the discrimination of nasality, we decompose the original signal using the EMD technique. Following this decomposition we extract the three first IMFs. This choice has been the subject of previous work [3].

Once these IMFs obtained, a sum of these, noted IMF123, is computed. An MFCC parameterization with 13 coefficients is applied to the signal extracted from the previous step.

For the classification stage, a multi layer neural network (ANN) is chosen. The input layer corresponds classically to the components of the observation vector (i.e. 13 inputs). The network has a single hidden layer and two outputs N/V. A Gaussian activation function is chosen for all the units. The training of the neural network is performed by the gradient backpropagation algorithm. The choice of this type of classifier is justified in [3]. The obtained results will be compared against an MFCC parameterization with 13 coefficients on the original signal.

For the extraction phase of the most discriminative IMF, we will use the signals obtained separately by each IMF (noted: IMF1, IMF2 and IMF3). In addition, the partial sums of these signals (i.e. IMF1 + IMF2 noted IMF12, IMF2 + IMF3 noted IMF23 and IMF1 + IMF3 noted IMF213) will be computed. For each signal, we apply an MFCC parameterization with 13 coefficients. Then we will use a neural network with the same architecture raised above for the classification stage.

Following the extraction of the signal allowing better classification N/V. We will try to increase discriminative power by using fusion by maximization [8] of the output of the obtained ANNs.
Table 1. Database distribution of each nasal vowels and each oral vowels for classification stage.

<table>
<thead>
<tr>
<th>Sons</th>
<th>Test</th>
<th>Train</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>For each vowel phone</td>
<td>1120</td>
<td>3360</td>
<td>4480</td>
</tr>
<tr>
<td>For each nasalized phone</td>
<td>1120</td>
<td>3360</td>
<td>4480</td>
</tr>
</tbody>
</table>

Table 2. Recognition rate of the classification N/V using an ANN classifier on the case of IMF123 and on the original signal with an MFCC parametrization with 13 coefficients.

<table>
<thead>
<tr>
<th>Tests</th>
<th>N</th>
<th>V</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMF123+MFCC</td>
<td>57.88%</td>
<td>92.12%</td>
<td>84.22%</td>
</tr>
<tr>
<td>Signal+MFCC</td>
<td>56.52%</td>
<td>90.83%</td>
<td>82.91%</td>
</tr>
</tbody>
</table>

3.2 Results

As we have considered a classification stage that implies a supervised training, we have decomposed the database in two parts: 43680 units for the training stage and 14560 units for the test stage. Table 1 shows their repartition between each nasal and oral vowels. For each phonetic unit, only the central part of 30ms, called frame, is processed.

The result presented in Table 2 shows that our method applied to IMF123 allows an improvement of classification N/V (i.e. 84.22%) compared with the standard one (i.e. 82.91%). This result is some what expected because we have just applied an MFCC to the original signal where we removed some low frequencies. Actually parts at low frequencies that we subtract to the original signal admit an almost negligible contribution in the total energy of this one. On the other hand, this observation confirms ones again the result obtained in our previous work [3]: all useful information for distinguishing between oral vowels and nasal vowels are in the three first IMFs.

We attemp now to answer the question: Can one say whether there is an IMF among these three which contains more information than the others or that contains the most discriminant information? To do this, we realized an MFCC analysis with 13 coefficients on each IMFs and an each partial summation of them. The results obtained are presented in table 3.

Initially, if we consider separately each IMF, we note that overall there is no visible improvement of recognition on all of the recognition rate. If now we consider the recognition rate observed in relation to nasalized phones, it is clear that the 3rd IMF allows over the two others the most significant reduction of the error.

Considering now the partial sums of individual IMFs issued from the EMD decomposition. The total recognition rate of classification approximately equal to the sum of the three first IMFs (i.e. IMF123) in the case of the partial sum IMF23. For cons, the addition of the IMF2 or IMF3 to the IMF1 allows some improvement. The latter is more significant if we consider the partial sum IMF13.

In conclusion, the signal composed of the partial sum IMF23 allows an lower error rate than the one obtained with the sum of the three first IMFs. However, by comparing the recognition rate in the nasal vowels (N) and the oral vowels (V), there is a significant improvement in terms of recognition rate on the oral vowels and a reduction on the detection of the nasalized one.

Now, a good idea would be to take advantage from the contribution of the partial sum IMF23 to improve the overall classification. To do so, we make recourse to a fusion by maximizing in the output of the classifier. This fusion will allow us to conclude on the contribution of each IMFs in the distinction between V and N. Table 4 shows the results of this fusion applied to each output of ANNs in the case of: the sum of IMF123 with IMF1, IMF2 and IMF3 and with the partial sum IMF12, IMF23 and IMF13.

Considering the results obtained, we note that following the fusion stage of IMF123 with each IMF, the IMF3 achieves the best compromise for the recognition rate. In the case of the partial sums, IMF23 allows a significant improvement of recognition rate by the fusion with IMF123.

To confirm the interest of IMF23, we conducted a fusion by maximizing between the partial sums. The table 5 shows the obtained results.

From another point of view, we know that: if we increase the number of coefficients derived from the MFCC,
Table 5. Recognition rate of the classification N/V using a fusion by maximinization in the output of each ANNs, in the partial summation.

<table>
<thead>
<tr>
<th>Tests</th>
<th>N</th>
<th>V</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMF12 &amp; IMF13</td>
<td>48.34%</td>
<td>95.14%</td>
<td>84.34%</td>
</tr>
<tr>
<td>IMF12 &amp; IMF23</td>
<td>51.23%</td>
<td>95.02%</td>
<td>84.92%</td>
</tr>
<tr>
<td>IMF13 &amp; IMF23</td>
<td>57.18%</td>
<td>94.07%</td>
<td>85.56%</td>
</tr>
</tbody>
</table>

Table 6. Recognition rate of the classification N/V using an ANN classifier on the case of IMF123 and on the original signal with an MFCC parametrization with 24 coefficients.

<table>
<thead>
<tr>
<th>Tests</th>
<th>N</th>
<th>V</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMF123+MFCC</td>
<td>67.92%</td>
<td>92.70%</td>
<td>86.98%</td>
</tr>
<tr>
<td>Signal+MFCC</td>
<td>66.44%</td>
<td>92.56%</td>
<td>86.54%</td>
</tr>
</tbody>
</table>

we obtain a better classification N/V. Thus, to confirm the robustness of our method, we conducted the same tests with 24 MFCC coefficients. Results are presented in table 6. Finally, we realize a fusion by maximization between IMF123 and IMF23 in the case of parametrization with 24 MFCC coefficients. Table 7 contain the obtained result. These tests clearly confirms the robustness of our technique.

4 Conclusion

The work [9] allows to conclude that the MFCC parametrization with 13 or 24 coefficients, in comparison with other existing techniques (e.g. LPCC, PLP) gives the best results in classification of nasal vowels vs. oral vowels.

In this work, we have concluded that the combined use of EMD with an MFCC parameterization allows a better discrimination of the characters of nasality present in the nasal vowels compared to the use of an MFCC parameterization on the original signal. Moreover, this study has clarified that the $3^{rd}$ IMF is the most discriminant for the detection of the character of nasality. This result joins these obtained by [10] that indicate that the coupling between the oral and nasal cavity induced spectral changes through the creation of formants and anti-formants in the low frequencies range. The originality of this work is incremental because it is the first in the area that deals with the EMD and nasality.

In our future work, we will try to improve the current system by improving the EMD decomposition using [11]. This will enable us to better frequency localization due to the elimination of edge effects.

References

[4] M. Saidi, O. Pietquin, R. André-Obrecht, EMD decomposition to discriminate nasal vs. oral vowels in French, Int. conf. on Speech and Computer (SPECOM 2009), St Petersburg, Russia, 2009, 313-318.
Figure 1. An example of EMD decomposition applied on a sum of cosines. (A) and (B) represent respectively the first and ninth iteration of the first sifting phase (j index in the algorithm). Dashed lines present EnvMin and EnvMax; the bold solid line shows the corresponding mean curve and the solid line is the original signal. In A, the second curve corresponds to h₁ and the third one to r₁ if h₁ is considered as the first IMF, before a new iteration of the sifting phase.

Figure 2. Illustration of the EMD decomposition of a signal containing three pure frequencies: 50Hz, 500Hz and 1KHz. From top to bottom: original signal, flowed by the corresponding IMFs (three IMFs issued from the application of the algorithm of the EMD decomposition) and the last one represents the residue of the decomposition.