Detection of Articulation Disorders using Empirical Mode Decomposition and Neural Networks

George Georgoulas, Voula C. Georgopoulos, *Senior Member, IEEE,* George D. Stylios and Chrysostomos D. Stylios, *Member, IEEE*

***Abstract*— This paper introduces a novel approach based on signal processing methods to extract features from speech signals and based on them to detect a specific type of articulation disorders. Articulation, in effect, is the specific and characteristic way that an individual produces the speech sounds. Empirical Mode Decomposition and the Hilbert Huang transform are applied to the speech signal in order to calculate the marginal spectrum of the signal. The marginal spectrum is subsequently subject to a mel-cepstrum like processing to extract features which are fed to a neural network classifier responsible for the identification of the articulation disorder. Our preliminary results suggest that this approach is very promising for the detection of the disorder under study.**

I. Introduction

Articulation refers to the actual production process of speech sounds in isolation or in words. It describes the process by which sounds, syllables, and words are formed when the tongue, jaw, teeth, lips, and palate alter the airflow in the vocal tract above the larynx. Articulation is a complicated procedure and often can be difficult to master since the ability to articulate speech begins in infancy and is a gradual process that continues through a child's sixth or seventh year. Due to the fact that the correct production of speech depends on many different physiological factors, articulation problems may frequently occur. According to the American Speech and Hearing Association [1] an articulation impairment is the atypical production of speech sounds that may interfere with intelligibility. Articulation errors can occur at the beginning, middle, or end of a word. They can affect both children and adults, and errors may range from a mild lisp to nearly unintelligible speech. Articulation errors are characterized by the omission, distortion, substitution, addition and/or incorrect sequencing of speech sounds [2] with the three first errors being the most common.

In a typical substitution error, for example, a child may say /θ/ instead of /s/ in the Greek word /salata/ (salad) so it would be heard as /θalata/. Another case is the omission error where the first syllable of the word may be omitted leaving only /lata/. These kinds of mistakes are systematic, which means that an individual may only misarticulate a couple of sounds, but they do so in all words that contain those sounds. In many cases, that results in unintelligible speech while in other cases the speech remains intelligible which is a fact that depends on the frequency of the misarticulated sounds. In any of these cases, the articulation disorder constitutes a problem for the patient that must be solved.

From the clinical practice and experience [3], a few of the most common substitution articulation errors that Greek children make are shown in Table I.

TABLE I.
SOME OF THE COMMON SUBSTITUTION ARTICULATION ERRORS IN GREEK

Target sound	Produced sound
/ɾ/	/γ/
/s/	/ʃ/, /θ/, or /ç/
/v/	/f/
/ð/	/θ/

The area of speech processing is one of the most active areas of signal processing and much work has been done for event detection in speech signals [4]-[5].

In this research work we investigate the use of a novel signal processing technique that analyzes the speech signal in an attempt to find a characteristic footprint for each one of the aforementioned articulation disorders.

Most real life processes are inheritably nonlinear and nonstationary. As a result, using techniques that assume linearity and stationarity, even though they are built on top of solid mathematical background, can be suboptimal, misleading or even have completely no connection to the physical system that they are supposed to "model". Empirical Mode Decomposition (EMD) and the Hilbert Huang transform (HHT), introduced in [6] came to fill this gap between theory and real life.

EMD lacks rigorous mathematical analysis and it decomposes the signal into a collection of Intrinsic Mode Functions (IMFs), where an IMF represents a simple oscillatory function with a number of conditions that have to

G. Georgoulas is with the Dept of Computer Applications in Finance and Management, TEI of Ionian Islands, Lefkas, Greece (e-mail: georgoul@teiion.gr).

V. Georgopoulos is with the Dept. of Speech and Language Therapy, TEI of Patras, Koukouli Patras, Greece (email: voula@teipat.gr).

G. D. Stylios is with the Dept of Computer Applications in Finance and Management, TEI of Ionian Islands, Lefkas, Greece (e-mail: gstylios@teiion.gr).

C. D. Stylios is with the Laboratory of Knowledge and Intelligent Computing, Dept. of Informatics and Communications Technology, TEI of Epirus, 47100 Artas, Kostakioi, Greece (email: stylios@teiep.gr)

978-1-4244-3553-1/09/$25.00 ©2009 IEEE

be satisfied. The well behaved Hilbert transforms of the IMFs give an alternative approach to time-frequency decomposition which results from the traditional short time Fourier transform and the wavelet transform [6].

In this research work, we investigate a combination of EMD and HHT methods as a novel hybrid approach to analyze the speech signal in order to extract a set of features capable of characterizing a specific misarticulation in the Greek language. Those features are the inputs to a multilayer perceptron (MLP) that performs the final stage classification.

The remainder of this paper is organized as follows: Section II describes the proposed methodology, briefly describing the combination of the EMD algorithm and the HHT, so as to extract suitable features that will feed the artificial neural network (ANN) classifier. In Section III, the results of the proposed approach to detect normally and misarticulated phonemes are presented and, finally, conclusions and future directions are included in Section IV

II. Materials and Methods

The overall approach is summarized in Fig 1. The approach can be divided into 5 stages. In the first stage all speech signals are normalized to have unit energy. Apart from that no further preprocessing has been applied. In the second stage a "frequency" spectrum is estimated using a newly proposed approach, based on the combination of EMD and Hilbert spectrum for non-linear and non-stationary time series analysis [6]. In stage 3 the calculated spectrum in the previous stage, is processed followed by a path similar to the one followed to extract the Mel-frequency cepstral coefficients (MFCCs). Then at stage 4, these coefficients/features are further processed using Principal Component Analysis (PCA) to uncorrelate them and also reduce the dimension of the feature vector. Finally stage 5, is the classification stage, where the previously extracted and processed features are eventually fed into an ANN that performs the detection

Each one of these five stages will be further described in the rest of this section along with a description of the data employed in this study.

A. Data Set

Our data set consists of samples of 16-bit precision, sampled at a rate of 44.1 KHz, which were collected from 144 children, ages 6-10, whose mother tongue was Greek. All children were asked to produce the pseudoword /asa/ three times each. Speech therapists were used as experts to evaluate and categorize the articulation of every child. Of the 144 children, 36 had normal production of the pseudoword. The rest 108 children produced articulation errors: 36 of them had substituted /s/ with /ʃ/, 36 of them had substituted /s/ with /θ/ and 36 had substituted /s/ with /ç/. As mentioned above, each child produced the target pseudoword three times and as a result we had 108 pseudowords for each one of the four classes.

The subjects' articulation tasks were recorded in a quiet room using a digital walkman Mini-Disc recorder, SONY® MZ-NH 1, with a SONY® ECM-MS907 unidirectional microphone, located at a 10 cm distance from the speaker's mouth. The files were of type .oma and were stored on a SONY 1GB Hi-MD minidisc. Then the files were transferred to a PC using SONIC STAGE version 4.0 and subsequently converted to type .wav using HMD Reader version 0.54.

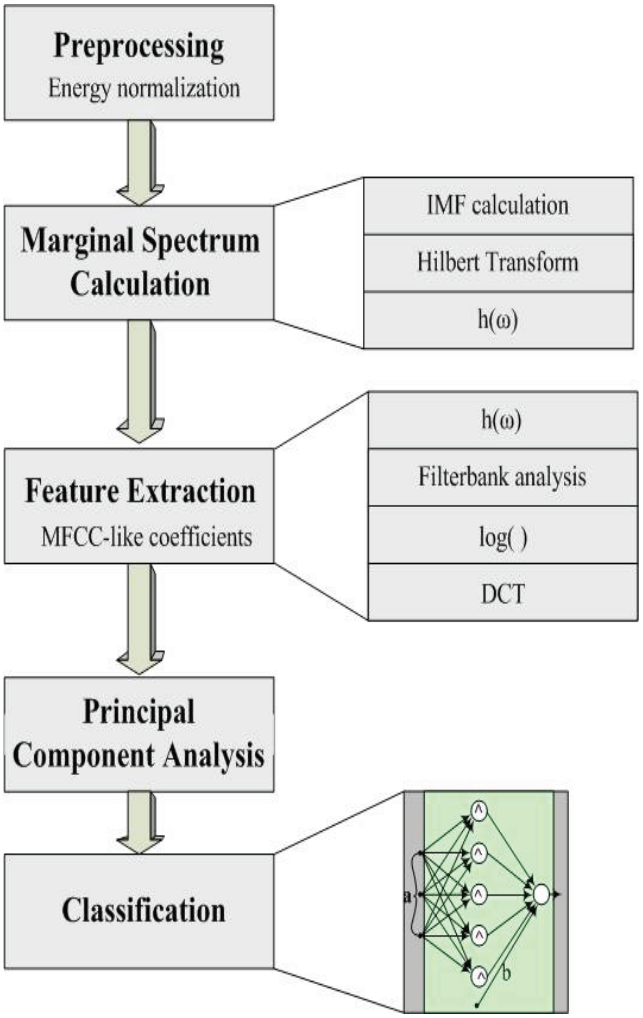

Fig. 1. The overall proposed methodology for the detection of articulation disorders.

B. Empirical Mode Decomposition

EMD is an algorithm that decomposes a signal into a finite set of oscillatory components (IMFs). These functions are symmetric with respect to a local zero mean and have the same number of zero crossings and extrema. The method for computing these functions was originally introduced by Huang et al. [6] and is implemented through the following 5 steps:

1. identify all minima and maxima of the given signal ($x(t)$)
2. create an upper ($e_{max}(t)$) and a lower ($e_{min}(t)$) envelope interpolating between successive maxima and minima respectively (usually via cubic interpolation)

3. calculate the running mean $m(t) = \frac{e_{\min}(t) + e_{\max}(t)}{2}$
4. subtract the mean from the signal to extract the detail $d(t)=x(t)-m(t)$.
5. repeat the whole process replacing $x(t)$ with $m(t)$ until the final residual is a monotonic function (or a user specific number of IMFs has been extracted – application dependant).

In practice, step 4 may not produce a valid IMF. As a result sifting needs to take place which implies the iteration of steps 1 to 4 upon the detail $d(t)$ until this fulfils the criteria of an IMF. Therefore the original signal $x(t)$ is eventually decomposed into a sum of IMFs plus a residual term:

$$x(t) = \sum_i IMF_i(t) + r(t) \quad (1)$$

as it is shown in Fig.2.

Following the implementation of the EMD algorithm, the Hilbert transform can be applied to each IMF separately and then the instantaneous frequency can be calculated as the derivative of the phase function.

After performing the Hilbert transform to each IMF the original signal can be expressed as the real part, RP, in the following form

$$x(t) = RP\left(\sum_j a_j(t)e^{i\theta_j(t)}\right) = RP\left(\sum_j a_j(t)e^{i\int \omega_j(t)dt}\right) \quad (2)$$

Equation 2 gives both the amplitude and the frequency of each component as a function of time. This time-frequency distribution of the amplitude is called the Hilbert-Huang spectrum ($H(\omega,t)$).

Then, by integrating over time we calculate the marginal spectrum:

$$h(\omega) = \int_0^T H(\omega,t)\,dt \quad (3)$$

The marginal spectrum offers a measure of total amplitude (or energy) contribution from each frequency value.

The frequency in either $H(\omega,t)$ or $h(\omega)$, has a totally different meaning from the Fourier spectral analysis [6]. While in the classical Fourier representation the existence of energy at frequency,ω means a component of a sine or a cosine wave persisted through the whole time span, in the case of the marginal spectrum the existence of energy at the frequency ω means only that in the whole time span, there is a higher likelihood for such a wave to have appeared locally.

Fig. 3 depicts the marginal spectrum of the pseudoword /asa/ spoken by a normal speaker and then Fig. 4 to Fig. 6 shows the marginal spectrum of the same word spoken by speakers with articulation disorders. As it can be seen and was pointed out in our previous work [7] the marginal spectrum could be used to extract relevant features for the detection of this specific articulation disorder.

The implementation of the EMD has been performed using the freely available MATLAB toolbox by Rilling et al. [8], [9] along with the TFTB toolbox developed by the same group [10].

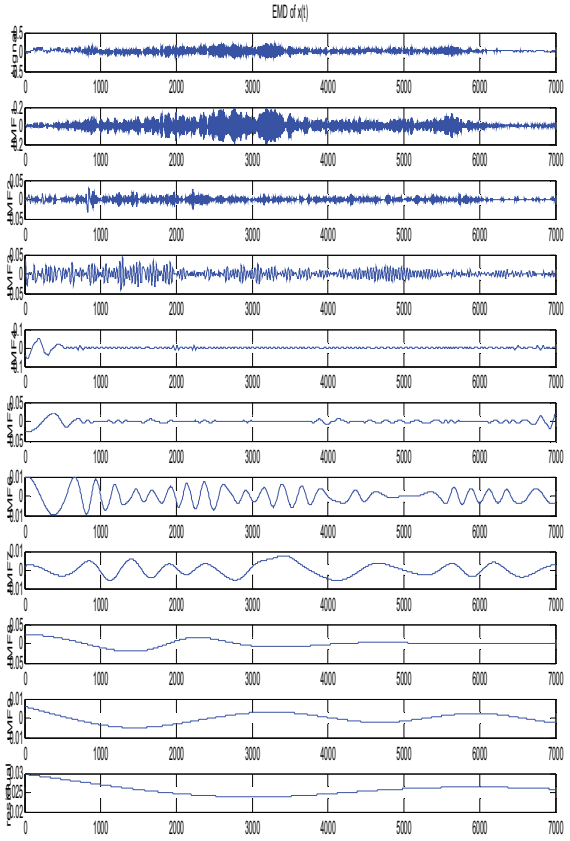

Fig. 2. Application of the EMD algorithm to phoneme /s/ produced by a normal speaker

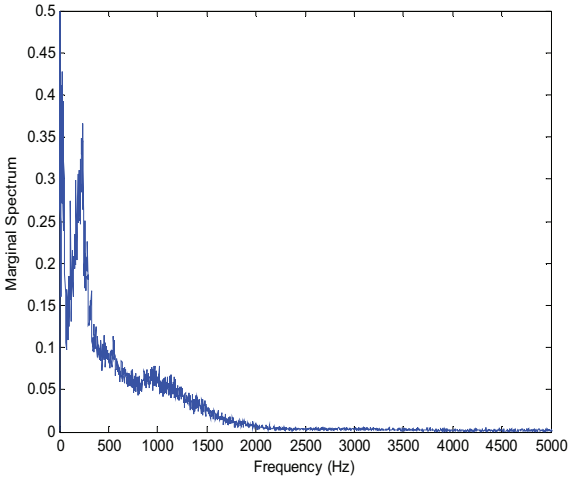

Fig. 3. Marginal spectrum of the pseudoword /asa/ spoken by a normal speaker.

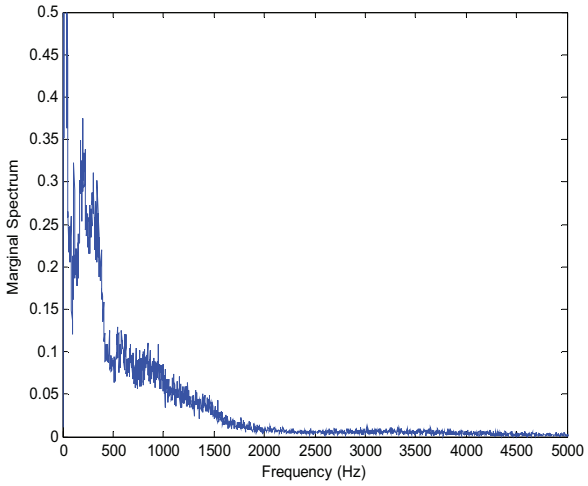

Fig. 4. Marginal spectrum of the pseudoword /asa/ spoken by a speaker with articulation disorder (/ʃ/).

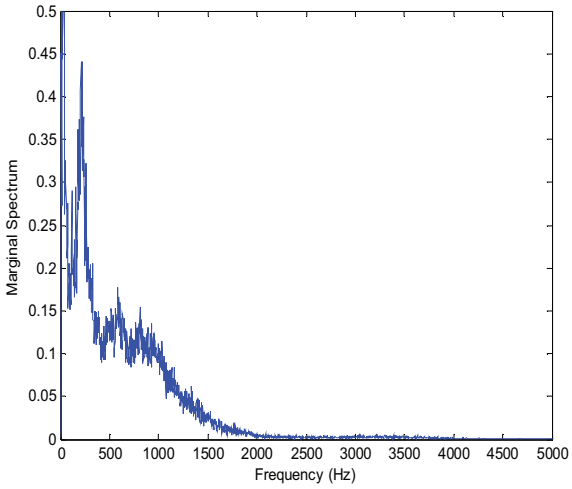

Fig. 5. Marginal spectrum of the pseudoword /asa/ spoken by a speaker with articulation disorder (/θ/).

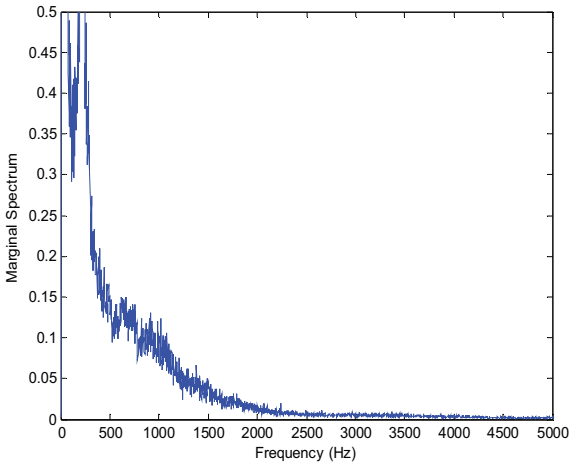

Fig. 6. Marginal spectrum of the pseudoword /asa/ spoken by a speaker with articulation disorder (/ç/).

C. *Multilayer Perceptron (MLP)*

ANNs are increasingly used in problem domains involving classification. They are networks composed of many simple processing elements, that operate in parallel and whose function is determined by the network's structure, the strength of their connection and the processing carried out by the processing elements (artificial neurons).

They are capable of finding commonalities in a set of seemingly unrelated data and for this reason are used in a growing number of classification tasks.

Among the numerous ANN paradigms encountered in the literature [11], the MLP is the most widely used in the field of pattern recognition [11]-[13]. Training of an MLP is often formulated as the minimization of an error function, such as the total mean square error between the actual output and the desired output summed over all available data. While the sum-of-squares error function is appropriate for regression, for classification problems it is often advantageous [13], [14] to optimize the network using the cross entropy error function (eq. 4 where, N is the number of training samples and c the number of classes) i.e. optimizing the network to represent the posterior probabilities of each class [12], [13].

$$E = -\sum_{n=1}^{N}\sum_{k=1}^{c}\left\{t_k^n \ln y_k^n + \left(1-t_k^n\right)\ln\left(1-y_k^n\right)\right\} \tag{4}$$

where $t_k^n \in \{0,1\}$ is a binary class label, (k=1,…,c) of the n^{th} data sample and y_k^n is the actual output of the k^{th} neuron of the ANN, when the n^{th} data sample is presented at its input.

In this case, we also need to use logistic activation functions for the hidden layer units and softmax (eq. 5) activation function for the outputs [12], [13].

$$y_j = \frac{\exp\left(a_j\right)}{\sum_i \exp\left(a_i\right)} \tag{5}$$

where a_i is the intermediate linear output of an artificial neuron.

The above configuration has proven to be more appropriate for classification purposes with many successful implementations [12], [13]. Therefore, in this research work the above formulation has been adopted.

D. *Feature Extraction*

In all pattern recognition problems the selection of an appropriate set of features is of paramount importance. The field of acoustic speech processing is not an exception to that [15]. Many different representations have been proposed among which the mel-cepstrum representation is probably the one most often used [16], [17].

The MFCCs are calculated in the following manner:

- A sound segment of length N is analyzed using the DFT.
- A filterbank is calculated (Fig. 7).
- The DFT power spectrum is multiplied with the (usually triangular) mel weighted filterbank.
- The result is summed to give the logarithmic mel

spectrum:

$$S(m) = \ln\left[\sum_{k=0}^{K-1} |X(k)|^2 H_m(k)\right] \quad (6)$$

where, K is the length of the DFT, $|X(k)|^2$ is the periodogram and $H_m(k)$ is the m^{th} (triangular) filter.

- We take the Discrete Cosine Transform of $S(m)$ which consist of the MFCCs:

$$c(q) = \sum_{m=0}^{M-1} S(m)\cos\left(q(m-0.5)\pi / M\right), q=1,...,Q \quad (7)$$

where Q is the number of cepstral coefficients and M the number of the filters in the filter bank.

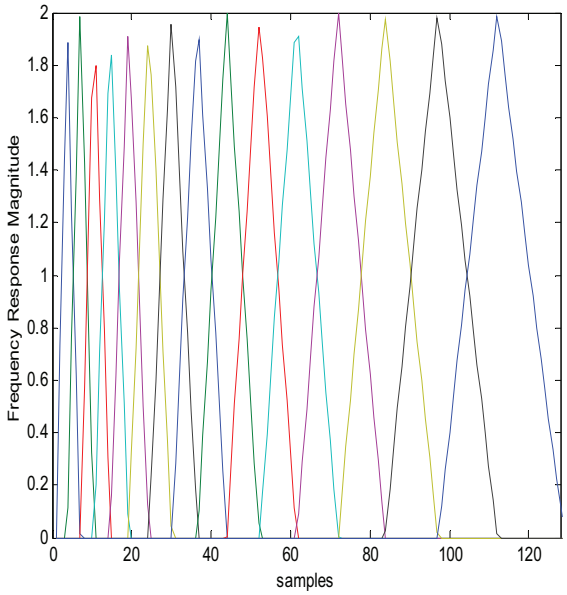

Fig. 7. Triangular filter bank created using the VOIBOX toolbox [18].

In the proposed approach, we have replaced the stage that calculates the DFT coefficients with the use of the marginal spectrum (eq. 3) coming from the HHT. This way, we produced MFCCs-like coefficients that we subsequently used as inputs to the MLP.

More specifically, we extracted 14 coefficients (excluding the zeroth coefficient as in the case of the conventional MFCCs) using a filterbank with 27 filters spanning one quarter of the sampling frequency (high end of highest filter equal to 0.25*44100 Hz).

E. Dimensionality Reduction Stage

In pattern recognition tasks, usually potential improvement (better generalization) can be achieved by using fewer features than those available [12]. Actually, when we build a classifier we tend to extract several features, which may convey redundant information about the pattern-class of interest. Therefore, in the proposed approach we included a PCA to uncorrelate the originally extracted features using a linear transformation [12].

PCA, or Karhunen-Loeve transformation, is a known approach to perform dimensionality reduction by linear combination of the original features in such a way that preserves as much of the relevant information as possible [11], [12]. This method computes eigenvalues of the correlation matrix of the input data vector and then projects the data orthogonally onto the subspace spanned by the eigenvectors (principal components) corresponding to the dominant eigenvalues. Even if the whole set of the eigenvectors is retained, this may also lead to an improvement of the classification performance, because the new set has features that are uncorrelated and this, in general, improves the classification capabilities of a classifier.

III. Results

In order to test the performance of our method we employed the stratified 9-fold cross validation scheme [19], which is an extension of regular multifold cross validation [11]. Therefore, we divided the 144 cases (subjects) into nine non-overlapping groups containing 16 cases each (four cases (subjects) from each class). For each one of the nine subsets, we created one training set comprising of the rest eight sets. Within the training set, we used 8-fold stratified cross validation to select the number of neurons in the hidden layer of the MLP and also select the number of retained principal components.

Once the parameters were optimized the MLP was retrained for this set of parameters using the whole training set and its performance was evaluated using the corresponding subset that was originally left out. The aforementioned procedure was adopted in order not to use the same set for both tuning and estimating the performance of the proposed classifier [20]. The mean accuracy and the mean confusion matrix were calculated as the average over the nine folds.

The results of our approach are summarized in Table II where the mean confusion matrix is depicted. The mean overall accuracy achieved was 79.86%.

TABLE II.
MEAN CONFUSION MATRIX

		Predicted class			
		/s/	/ʃ/	/θ/	/ç/
True class	/s/	83.33	11.11	1.85	3.70
	/ʃ/	11.11	77.78	1.85	9.26
	/θ/	5.56	0.93	77.78	15.74
	/ç/	3.70	5.56	10.19	80.56

IV. Conclusions and Future Work

In this research work we proposed a novel integrated method for the detection of a common substitution articulation error in Greek language. The proposed methodology seems to perform reasonably well, but there are still certain issues that have to be considered. This is the first time -to the best of

our knowledge- that the marginal spectrum coming from the HHT is used along with this mel-scale approach in order to extract features. In our preliminary experiments where a linear division of the frequency band was attempted, the mean classification accuracy (not show in this paper) reached a value of around 65%. This alone pinpoints the merit of using the logarithmic approach of the MFCCs. In future work we will try different classifiers – Support Vector Machiness, Radial Basis Function Neural Networks - in order to see if better accuracy can be achieved.

From Table II, which presents the classification results, we can conclude that the proposed approach performs pretty well in discriminating between the four different classes. Moreover it seems that there is quite an overlap between the /s/ and /ʃ/ classes on one hand and the /θ/ and /ç/ on the other hand. This means that we might need to built upon a modular/hierarchical approach following the "divide and conquer" principal.

It is known that the shape, the bandwidth and number of bands plays an important role in the feature extraction process. In this work we didn't try to find an optimal set of the aforementioned design parameters. In future work we will investigate and identify those bands that are more sensitive in picking the difference stemming from the different types of articulation disorders in a wrapper approach. In doing so we will employ different types of kernels (Gaussians, triangulars etc.) shifted and scaled using an evolutionary approach.

Moreover even though this approach seems quite promising for this speaker independent recognition problem, it doesn't explore the time-frequency capabilities of the HHT. More specifically the time information was not retained leaving room for potential improvement.

In future work we will also test our method with other types of articulation disorders to check its validity as an alternative to the classic mel-cepstrum approach.

Finally for a fully automated system for the detection of articulation disorders an auto-segmentation stage is required to replace the cumbersome and non practical manual preprocessing of the speech signal.

REFERENCES

[1] American Speech-Language-Hearing Association Ad Hoc Committee on Service Delivery in the Schools. Definitions of Communication Disorders and Variations. ASHA, 35 (Suppl. 10), 40-41, 1993

[2] Bauman-Waengler, J. *Articulatory and Phonological Impairments: A Clinical Focus*. Boston: Allyn & Bacon, 2000

[3] V. C. Georgopoulos, G. A. Malandraki, and C. D. Stylios, "A computer based speech therapy system for articulation disorders," *in Proc. 4th Int. Conf. Neural Networks and Expert Systems in Medicine and Healthcare (NNESMED)*, Milos Island, Greece, pp. 223-230, 2001

[4] B. Yegnanarayana and R. N. J. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 4, pp. 313-327, July 1998.

[5] H. Kawahara, Y. Atake, and P. Zolfaghari, "Accurate Vocal Event Detection Method based on a Fixed-point to Weighted Average Group Delay," *in Proc. of ICSLP'2000*, Beijing, China, 2000, vol. IV, pp. 664-667.

[6] N. E. Huang, Z. Shen, S. R. Long, M. L.Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu (1998) The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis, *Proceedings of the Royal Society London A*, Volume 454, pp. 903–995

[7] G. Georgoulas, V.C. Georgopoulos and C. D. Stylios Investigating Articulation Disorders Using Empirical Mode Decomposition, ESBME 2008

[8] G. Rilling, P. Flandrin and P. Goncalves, On Empirical Mode Decomposition and its Algorithms, *in Proc. IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing, NSIP03*, Grado, Italy, 2003

[9] http://perso.ens-lyon.fr/patrick.flandrin/emd.html

[10] http://tftb.nongnu.org/

[11] S. Haykin, *Neural Networks: A Comprehensive Foundation*. 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 1999.

[12] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, New York, 1995

[13] R. A. Dunne, *A statistical Approach to Neural Networks for Pattern Recognition*, John Wiley & Sons, New Jersey, 2006

[14] P.Y. Simard, D. Steinkraus, and J.C. Platt, Best practices for convolutional neural networks applied to visual document analysis, *in Proceedings, Seventh International Conference on Document Analysis and Recognition*, 2003

[15] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th edition, Academic Press, 2008.

[16] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, N.J. 1993. Digital representation

[17] S. S, Davis and P. Mermelstein, Comparison of Parametric Representations for Monosylabic Word Recognition in Continuously Spoken Sentences, *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no, 4, 1980

[18] VOICEBOX: Speech processing Toolbox for Matlab, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[19] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont: Wadsworth International Group, 1984

[20] S. L. Salzberg, "On Comparing Classifiers: Pitfalls to avoid and a Recommended Approach," *Data Mining and Knowledge Discovery*, vol. 1, pp 317-328, 1997.