Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Speech Sound Classification and Detection of Articulation Disorders with Support Vector Machines and Wavelets

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

Abstract—**This paper proposes a novel integrated methodology to extract features and classify speech sounds with intent to detect the possible existence of a speech articulation disorder in a speaker. Articulation, in effect, is the specific and characteristic way that an individual produces the speech sounds. A methodology to process the speech signal, extract features and finally classify the signal and detect articulation problems in a speaker is presented. The use of Support Vector Machines (SVMs), for the classification of speech sounds and detection of articulation disorders is introduced. The proposed method is implemented on a data set where different sets of features and different schemes of SVMs are tested leading to satisfactory performance.**

I. Introduction

Articulation refers to the production process of speech sounds in isolation or in words. The process describes the physiological movements involved in modifying the airflow, in the vocal tract above the larynx, for the production of the various speech sounds. In essence sounds, syllables, and words are formed when the vocal chords, tongue, jaw, teeth, lips, and palate change the stream of air that is produced by the respiratory system. Articulation is a complicated procedure and often can be difficult to master. Due to the fact that the correct production of speech is dependent on many different physiological factors, articulation problems may frequently occur. They appear when a person produces sounds, syllables or words incorrectly so that listeners do not understand what is being said or need to pay more attention to the way the words sound than to what they mean. Articulation errors become most noticeable in the rapid production of sounds required in usual conversation. Articulation problems affect both children and adults, and errors may range from a mild lisp to nearly unintelligible speech. Most articulation errors fall into one of three categories: omissions, substitutions, or distortions.

G. Georgoulas is with Laboratory for Automation and Robotics, University of Patras, 26500 Patras, Greece (Corresponding author Tel +302610997293; Fax: +302610997309; Email: georgoul@ee.upatras.gr)

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

C. Stylios is with the Dept. of Communications, Informatics and Management, TEI of Epirus, Kostakioi, Artas, Greece (email: stylios@teiep.gr)

In a typical substitution error, for example, a child may say /θ/ instead of /s/ in the Greek word /sela/ (saddle) so it would be heard as /θela/. Another case is the omission error where the second syllable of the word may be omitted leaving only /se/. These kinds of mistakes are systematic, which means that a child may only misarticulate a couple of sounds, but he/she does 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 [1], 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/	/ɣ/
/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 [2]-[3]. In this research work we propose a method that analyzes the speech signal in order to extract appropriate features combined with an advanced learning paradigm from the field of pattern recognition for the discrimination of articulation disorders.

Support Vector Machines (SVMs) have gained great attention and have been used extensively and, most importantly, successfully in the field of pattern recognition [4]-[7]. Recent findings have shown that implementation of SVMs creates reliable classifiers (i.e. classifiers with very good generalization performance) even in high dimensional spaces and under small training sample conditions [8].

However, a more important stage before the classification is the representation and feature extraction from signals. The produced and selected features have to be suitable for recognition of phonemes and sounds. For this research work, features will be extracted using the discrete wavelet transform.

Wavelet analysis is a quite novel signal processing method. The way wavelet analysis localizes the signal's information in

the time-frequency (or time-scale), makes it especially suitable for the analysis of non-stationary signals and as an alternative to the classical short-time Fourier transform [9]. Figure 1 depicts the phoneme /s/ and the wavelet coefficients up to level 3 using Daubechies wavelets with 4 vanishing moments.

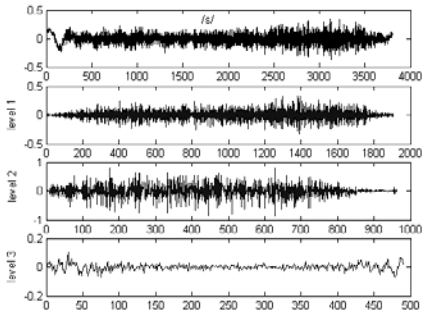

Fig. 1. The phoneme /s/ and the wavelet coefficients, up to level 3, using Daubechies wavelets with 4 vanishing moments.

The diagnosis and treatment of articulation disorders is very difficult and lies in the expertise of speech and language pathology. Treatment is critical if one considers the possible impact of an articulation problem on one's social, emotional, educational, and/or vocational status. Since speech is the most important means of communication, the quality of an individual's life is affected by the adequacy of it.

Children often mispronounce the Greek fricative sound /s/, as shown in Table I. In this research work we will examine this fricative sound as an example to identify articulation errors.

The remainder of this paper is organized as follows: in the following section the materials and methods are presented. First the data set used for analysis is described and then the methods of wavelets and SVMs are briefly presented. Next three feature sets based on the wavelet transform are shown. The results based on the SVM classifier for each of the three feature sets are demonstrated. Finally, conclusions and future directions are included.

II. Materials and Methods

A. Data Set

Samples were collected from 36 children ages 6-8 whose mother tongue was Greek. All children were asked to produce the pseudoword /asa/. Speech therapists were used as experts to evaluate and categorize the articulation of children. Of the 36 children 12 had normal production of the pseudoword, and 24 produced articulation errors of which 12 were substitution of /s/ with /ʃ/ and 12 were substitution of /s/ with /θ/.

The /s/, /ʃ/ and /θ/ sounds were isolated, truncated to the same time length and the energy was normalized to value 1.

B. Wavelets

In the past few years, wavelet analysis has been found to be particular useful in the field of biomedical signal processing [10], [11] as an alternative to short-time Fourier transform. The intrinsic property of the wavelet transform to localize well both in time and frequency domain makes it very appealing in case of nonstationary signals. Even for stationary signals, it can be sometimes difficult to choose a good resolution to analyze the signal. This is the case when the signal contains various features at different resolutions [12].

The continuous wavelet transform (CWT) of signal $s(t)$ is produced taking the inner product of the signal with translated and scaled versions of a (real or complex) analyzing function, also called mother wavelet ψ.

Translations and dilations of this "mother" (or analyzing) wavelet (Eq. 1) are used to transform the signal into another form (time-scale representation).

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \quad (1)$$

In the case of discrete parameter wavelet transform (DPWT) [13], the dilation and translation parameters α, b are restricted only to discrete values leading to the following expression:

$$\psi_{m,n}(t) = a_0^{-m/2}\psi\left(\frac{t-nb_0a_0^m}{a_0^m}\right) \quad (2)$$

The choice of $a_0 = 2$ and $b_0 = 1$ (dyadic grid arrangement) is quite usual

$$\psi_{m,n}(t) = 2^{-m/2}\psi\left(2^{-m}t-n\right) \quad (3)$$

For discrete time signals the discrete time wavelet transform (DTWT) [13] is given by:

$$T_{m,n} = 2^{-m/2}\sum_k x(t)\psi\left(2^{-m}k-n\right) \quad (4)$$

As it is obvious, different mother wavelets give rise to different classes of wavelets, and thus, the behavior of the decomposed signal can be quite different. In this work we experimented using Daubechies, coiflets and biorthogonal families, with different number of vanishing moments. All the aforementioned wavelets were developed by Daubechies [9] and they have the appealing property of having compact support and the wavelet transform can be computed with finite impulse response conjugate mirror filters using a fast filter bank algorithm.

C. Support Vector Machines

Support Vector Machines are learning systems that are trained using an algorithm from optimization theory. The aim of a support vector classifier is to "construct" a good separating hyperplane in a high dimensional feature space. More specifically, given a data set $D = \left\{\left(\mathbf{x}_i, y_i\right)\right\}_{i=1}^{n}$ of labeled

examples $y_i \in \{-1,1\}$ and a kernel function K, through an optimization process for each $\mathbf{x}_i$ a coefficient a_i such as to maximize the margin between the hyperplane and the closest instances to it is found. Every new pattern $\mathbf{x}$ is classified to either one of the two categories through

$$f(\mathbf{x}) = sign\left(\sum_{i=1}^{n} y_i a_i K(\mathbf{x}, \mathbf{x}_i) + b\right) \quad (5)$$

where b is a threshold parameter. The coefficients a_i are found by solving the following quadratic programming problem of maximizing:

$$\sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad (6)$$

subject to $0 \le a_i \le C, \quad i = 1,...,n$ and $\sum_{i=1}^{n} a_i y_i = 0$

In the above formulation (L1 Soft-Margin SVM) C is the margin parameter that determines the trade-off between the minimization of the classification error and the maximization of the margin. Depending on the choice of the kernel function different hyperplanes and different classifiers are constructed. Among the most popular kernels for classification tasks is the Gaussian Radial Basis Function kernel given by:

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{-\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right) \quad (7)$$

Even though SVMs were primarily designed for binary classification problems, they can also be used to deal with multi-class classification problems. The most common approaches to create M-class classifiers, are the "one versus the rest" and the "pairwise classification" [14]. Taking into account that those methods can give comparable results and generally no multi-class approach outperforms the others, we only considered the pairwise classification (or "one to one") approach. According to this approach, a classifier is trained for every possible pair of classes. That is, for a problem with M classes, it results in (M-1)M/2 binary classifiers (in our case M=3 and we have to train only 3 classifiers, the same number of classifiers as in the case of one versus the rest). After the training of the classifiers in order to classify a test pattern we evaluate the output of each one of the classifiers and the pattern is classified to the class that gets the highest number of "votes". To perform the experimental tests, we used the software package LIBSVM [15] to design and apply SVMs and for the wavelet analysis the wavelet toolbox of Matlab®.

III. Feature extraction

The most appealing characteristic of wavelets is that they have the ability to decompose a signal into a number of scales, where each scale represents a particular "coarseness" of the signal under study [16]. For this work we performed discrete-time wavelet transform up to level 5 using various mother wavelets and constructed three feature sets:

First feature set: For each scale we calculated the corresponding standard deviation of the distribution of wavelet coefficients ending to 5 features for each signal. This feature set quantifies the way that the "power" of the signal is segregated at different scales.

Second feature set: For each level, we calculated the corresponding entropy S of the (discrete) distribution of normalized energies of wavelet coefficients (i.e. squared magnitudes) - Shannon entropy measure-

$$S = -\sum_{i=1}^{N} p_i \log(p_i) \quad (8)$$

with $\sum_{i=1}^{N} p_i = 1$ and $p_i \log(p_i) = 0$, if $p_i = 0$

resulting again in 5 features for each signal. This feature set quantifies the concentration of coefficients at different scales (the more clustered the distribution of coefficients, the lower the entropy).

Third feature set: The third feature set simply consisted of the aggregation of features from the first and second feature sets (i.e. a total of 10 features for each signal).

The selection of the standard deviation of the wavelet coefficients at different scales, as well as the entropy of the normalized entropies, were chosen on the belief that the misarticulated sounds would result in a different distribution of the wavelet coefficients something which can be quantified by those 2 parameters.

IV. Results

Due to the small amount of data, in order to estimate the overall classification performance of the proposed method we employed the leave one out procedure [17] i.e. each time a SVM classifier was trained using 35 samples and tested on the remaining case. This procedure was repeated 36 times and the overall classification performance is the number of correct classifications on the single test case divided by 36. However, in the case of multi class problems, the overall classification rate cannot fully represent the performance of the classifier. A more descriptive sight can be given by a confusion matrix [18]. In Fig. 2-4 the confusion matrices of the best constructed classifier (in terms of overall classification performance) for each one of the 3 experimental setups is depicted.

		True class		
		/s/	/θ/	/ʃ/
Predicted class	/s/	8	0	2
	/θ/	2	12	2
	/ʃ/	2	0	8

Fig. 2. Confusion matrix for the SVM with the best overall classification performance for the first experimental setup.

		True class		
		/s/	/θ/	/ʃ/
Predicted class	/s/	12	2	0
	/θ/	0	10	2
	/ʃ/	0	0	10

Fig. 3. Confusion matrix for the SVM with the best overall classification performance for the second experimental setup.

		True class		
		/s/	/θ/	/ʃ/
Predicted class	/s/	8	0	0
	/θ/	2	12	2
	/ʃ/	2	0	10

Fig.4. Confusion matrix for the SVM with the best overall classification performance for the third experimental setup.

As it can be seen, the best overall classification performance (88.89%) was achieved using the second feature set, where features are the entropy of the normalized coefficients for levels 1 to 5 using a biorthogonal spline wavelet (with 8 vanishing moments for the decomposition and 6 for the reconstruction wavelet). Tthis feature set identifies correctly (100%) the non-erroneous articulation /s/, which is very important for the diagnosis of articulation disorders.

For comparison reasons we also tested the performance of "conventional" classifiers with the set of features achieving the best performance when combined with an SVM classifier (i.e. the second feature set). To be more specific, we tested the performance of the linear and the quadratic discriminant classifiers [18]. The best results are summarized in Fig. 5 and were achieved using a linear classifier and features extracted using symmlet wavelets. The overall classification performance is 77.78%. Compared to the SVM classification scheme the linear discriminant classifier (ldc) fails to categorize correctly all the normal pronunciations.

		True class		
		/s/	/θ/	/ʃ/
Predicted class	/s/	8	2	2
	/θ/	4	10	0
	/ʃ/	0	0	10

Fig.5. Confusion matrix for the ldc with the best overall classification performance for the second experimental setup.

V. Conclusions-Discussion

The proposed methodology seems to perform quite well, but there are still some issues that have to be considered. First of all in this work we didn't include any feature selection or feature reduction stage. Moreover we included all the 5 levels without investigating if some of them are more informative concerning the reflection of the particular disorders.

As it is shown the use of ten features does not improve the classification performance. This enhances the belief that in future work we should include a feature selection stage in order to keep all the relevant information and eliminate any redundancy present among the different features.

Furthermore, certain other capabilities are offered by the wavelet transform that can be utilized in future work. By using global statistics we exploit only the scale property of the wavelet transform and we do not take into account the evolution of this non-stationary phenomenon. In future work we will concentrate more thoroughly into this matter.

Even though the proposed method is promising, it still has to be tested using a bigger data set with words, pseudowords, and conversation before safer conclusions can be drawn.

Acknowledgment

The Project is co-funded by the European Social Fund and National Resources - (EPEAEK-II) ARCHIMIDIS-I.

References

[1] 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, 2001, pp. 223-230,

[2] 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.

[3] 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.

[4] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.

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

[6] K. Veropoulos, N. Cristianini, and C. Campbell, "The Application of Support Vector Machines to Medical Decision Support: A Case Study," *Advanced Course in Artificial Intelligence (ACAI'99)*, July 1999.

[7] K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An Introduction to Kernel-Based Learning Algorithms," *IEEE Trans. Neural Networks*, vol. 2, no. 2, pp. 181-201, March 2001.

[8] H. Byun, and S. W. Lee, "Application of Support Vector Machines for Pattern Recognition: A Survey," in *Proc. 1st International Workshop SVM 2002*, LNCS vol. 2388, pp. 213-236.

[9] I. Daubechies, *Ten Lectures On Wavelets*. Philadelphia: Siam, 1992.

[10] M. Unser, and A. Aldroubi, "A Review Of Wavelets In Biomedical Applications," in *Proc. IEEE*, 1996, vol. 84, no. 4, pp. 626-638.

[11] G. Georgoulas, C. D. Stylios, and P. P. Groumpos, "Feature Extraction And Classification Of Fetal Heart Rate Using Wavelet Analysis And Support Vector Machines," *Int. J. Artificial Intelligence Tools*, vol. 15, no. 3, pp. 411-432, June 2006.

[12] M. Akay, *Detection And Estimation Methods For Biomedical Signals*. San Diego, CA: Academic Press, 1996.

[13] E. C. Ifeachor and B. W. Jervis, *Digital Signal Processing: a practical approach*. Edinburgh Gate: Pearson Education Limited, 2001.

[14] B. Scholkopf, and A. J. Smola, *Learning With Kernels. Support Vector Machines, Regularization, Optimization, And Beyond*. Cambridge, MA: MIT Press, 2002, pp. 211-214.

[15] C. C. Chang, and C. Lin, "LIBSVM: A Library For Support Vector Machines," 2003.

[16] S. Mallat, "A Theory For Multiresolution Signal Decomposition: The Wavelet Representation". *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, no. 7, pp. 674-793, 1989.

[17] S. M. Weiss, and C. A. Kulikowski, *Computer Systems That Learn*. San Mateo, CA: Morgan Kaufman, 1990.

[18] A. Webb, *Statistical Pattern Recognition.* UK: John Wiley & Sons, 2002.