DWT-BASED PHONETIC GROUPS CLASSIFICATION USING NEURAL NETWORKS

Pham Van Tuan and Gernot Kubin

Signal Processing and Speech Communication Laboratory
University of Technology, Graz, Austria
pvtuan@inw.tugraz.at ; g.kubin@ieee.org

ABSTRACT

This paper presents an improvement of the Discrete Wavelet Transform (DWT)-based phonetic classification algorithm by using Neural Networks (NN) to learn optimal thresholds for speech classification. Two feedforward NNs (two layers) operate on input features extracted from speech frames (10ms length) by DWT and statistical measurement in order to classify these frames as transient, voiced vowel, voiced consonant and unvoiced consonant categories. Hard thresholds in our earlier paper are used to detect silence and voiced closure intervals. The new algorithm is tested with the TIMIT database and compared with other algorithms to demonstrate its superior performance.

1. INTRODUCTION

The classification of speech frames is an important step in many speech processing applications. Some speech coding systems require phonetic classification to determine the optimal bit allocation for every different speech frame. The discrimination between phonetic classes will improve quality and performance of data-driver speech synthesizers. Speech classification has been studied by a variety of methods. Some linear speech classification such as [7] use statistical features (relative energy level, zero crossing rate,etc) of the speech signal to decide about voiced, unvoiced or silence. In other approaches, the specific spectral charactersitic of speech sounds are exploited by optimal filters designed to discriminate among speech classes [3]. Furthermore, some DWT-based algorithms have been developed to achieve classification at the phonetic level [1],[2]. Since the development of the backpropagation learning algorithm, the feedforward NN has been used widely in pattern recognition and, in particular, for speech classification with promising potential [4], [5], [6].

In this paper, we propose a new combined system of linear and NN-based nonlinear classifiers using wavelet parameters as well as statistical parameters to increase the performance and robustness of an optimal threshold-based speech classifier. The goal is to classify 10ms speech frames into phonetic categories [1]. Then, smoothing and interpolation techniques are used to mark boundaries between *phonetic groups* which we define as homogeneous sequences of speech sounds that belong to the five phonetic classes considered above. One of our new contribution is the consideration of so-called "wavelet-detail ratio features" across two neighbouring frames as input for the transient detection. Gender-dependent and gender-independent NN classifiers are built to study and evaluate the impact of speaker dependency on speech classification.

2. ADVANTAGES OF DWT IN SPEECH PROCESSING

Any signal $s(t)$ can be represented with basic functions as:

$$s(t) = \sum_m \sum_n d_{m,n}\psi_{m,n}(t) \tag{1}$$

where $\psi_{m,n}(t) = a_0^{-m/2}\psi(a_0^{-m}t - nb_0)$ are basis functions and $d_{m,n}$ are the coefficients of the DWT of $s(t)$ for all $m, n \in Z$,

$$d_{m,n} = a_0^{-m/2}\int_{-\infty}^{\infty} s(t)\psi(a_0^{-m} - nb_0)dt \tag{2}$$

With $a_0 = 2$ and $b_0 = 1$, we obtain the dyadic DWT. The advantage of DWT in speech processing is based on the relation between DWT and multiresolution analysis (MRA) which provides the structure for a multiscale decomposition of a signal. If we have a function $s(t)$, it can be decomposed into the sum of a low-pass approximation plus L details at L resolution stages as:

$$\begin{aligned} s(t) = &\sum_{n=-\infty}^{\infty} c_{L,n}2^{-L/2}\phi(\frac{t}{2L} - n) + \\ &\sum_{m=1}^{L}\sum_{n=-\infty}^{\infty} d_{m,n}2^{-m/2}\psi(\frac{t}{2^m} - n) \end{aligned} \tag{3}$$

where $\phi(t)$ and $\psi(t)$ are a scaling function and a wavelet function, $c_{m,n}$ and $d_{m,n}$ are the approximation or scaling coefficients (low-frequency part) and the detail or wavelet coefficients (high-frequency part) of the output of the DWT which are given by:

$$c_{m,n} = \sqrt{2}\sum_{k=-\infty}^{\infty} c_{m-1,k}h_0(k - 2n) \tag{4}$$

$$d_{m,n} = \sqrt{2}\sum_{k=-\infty}^{\infty} c_{m-1,k}h_1(k - 2n) \tag{5}$$

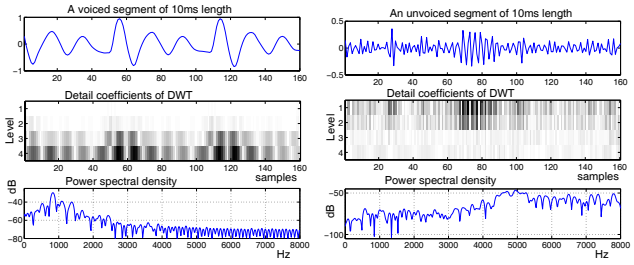
where $h_0(n)$ and $h_1(n)$ are synthesis low-pass and high-pass responses of a two-band paraunitary filter bank.

Typically, the energy of the voiced vowel frames is mostly contained in the approximation part and much less in the detail part and vice verse for the unvoiced consonant frames. The relative equal energy distribution occurs in the voiced consonant frames. So, this property can be used in the specific representation of the three different classes. Furthermore, the signal is decomposed into different scale levels, and the energy of the details varies over different scales in different ways, depending on the input signal (a

property similar to spectral tilt). We observe this energy variation of the detail coefficients for voiced vowel frames in Fig. 1-a and unvoiced consonant frames in Fig. 1-b as:

$$E(d_{1,n}) < E(d_{2,n}) < E(d_{3,n}) < E(d_{4,n}) \quad (6)$$

$$E(d_{1,n}) > E(d_{2,n}) > E(d_{3,n}) > E(d_{4,n}) \quad (7)$$



(a) voiced frame (b) unvoiced frame

Fig. 1: Energy variation of detail coefficients over 4 levels and power spectral density

From the observation at the first three levels of wavelet analysis in Fig. 2, we can also apply the energy variation of detail coefficients for transient classification when considering two neighboring frames. A transient frame always has higher absolute energy in its details coefficients than a closure interval frame which may be silence or periodic. This characteristic is used to define the closure interval-transient detail energy ratio and combining with other statistical features to detect transient frames.

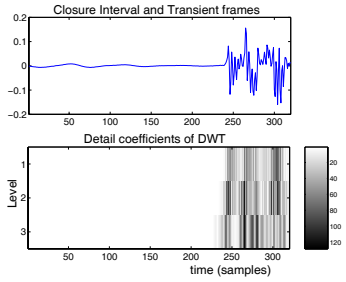


Fig. 2: Different energy variation of detail coefficients over first 3 levels for a closure interval followed by a transient frame.

3. FEATURE EXTRACTION

As discussed in the introduction, we want to classify five types of phonetic groups which are homogeneous frames sequences having the same phonetic characteristics as follows:

- A silence group are frames which have a very low overall amplitude level and a blank spectrogram.
- A voiced vowel group includes vowels, semivowels and diphthongs which have a repetitive time-domain structure and low-frequency voiced striations in the spectrogram.
- A voiced consonant group includes voiced and glottal fricatives which have both periodic and noise-like properties, and nasals, which have a weak and interrupted voice bar in the spectrogram.
- An unvoiced consonant group includes only unvoiced fricatives which have an irregular time-domain structure and only high frequencies in the spectrogram.
- A transient group include plosives and affricates which contain a transient frame inside.

As the basis for the phonetic classifier, we need to extract the following representative features from each frame which has 10ms length corresponding with K=160 samples:

- ***Approximation wavelet energy ratio (AWER)*** is the ratio of the energy in the approximation coefficients and the energy of all wavelet coefficients at the first level of wavelet analysis:

$$AWER = \frac{\sum_{n=1}^{N_1} (c_{1,n})^2}{\sum_{n=1}^{N_1} (c_{1,n})^2 + \sum_{n=1}^{N_1} (d_{1,n})^2} \quad (8)$$

  where $c_{1,n}$ and $d_{1,n}$ are the approximation and detail coefficients (length N_1) at the first level wavelet analysis.

- ***The energy variation of detail coefficient (EVD)***, (see Fig. 1) is defined by the following equation:

$$EVD(m,k) = \frac{1}{N_m} \sum_{n=1}^{N_m} (d_{m,n})^2 - \frac{1}{N_k} \sum_{n=1}^{N_k} (d_{k,n})^2 \quad (9)$$

  where N_m and N_k are lengths of detail coefficient sequences $d_{m,n}$ and $d_{k,n}$ at different analysis level m and k.

- ***The closure interval-transient detail ratio (CTDR)*** is the ratio of the detail coefficient energies at the same wavelet analysis level, computed for the closure interval and the following transient frame in Fig. 2:

$$CTDR(m) = \frac{\sum_{n=1}^{N_m} (d^2_{m,n})^2}{\sum_{n=1}^{N_m} (d^1_{m,n})^2} \quad (10)$$

  where $d^1_{m,n}$ and $d^2_{m,n}$ are the wavelet detail coefficients of the closure interval and the transient frame, N_m is the length of the sequences $d^1_{m,n}$ and $d^2_{m,n}$.

- ***Short-term average energy (SAE)*** is calculated for each frame:

$$SAE = \frac{1}{K} \cdot \sum_{i=1}^{K} \left(\frac{s(i)}{s_p}\right)^2 \quad (11)$$

  where $s(i)$ are the samples of 10ms frame, and s_p is the peak value of the input signal.

- ***Zero crossing rate (ZCR)*** is a measure of frequency content of each speech frame:

$$ZCR = \sum_{i=1}^{K} |sgn[s(i) - sgn(s(i-1))]| \quad (12)$$

4. NEURAL NETWORK CLASSIFIER

4.1. Network configuration setup

A supervised learning algorithm is used to train a two-layer feed-forward network. This means that the network weights and biases are adjusted to minimize the mean square error between the network outputs and real target outputs.
We train a first network with 5-dimensional input vectors and 1-dimensional output to detect transient frames. Second network is configured with 4-dimensional input vectors and 3-dimensional output to perform the three-way classification: voiced vowel frame, voiced consonant frame and unvoiced consonant frame. The output is labeled as 1 for the desired frames and 0 for other frames.

As a preprocessing step, all elements of the input and output vectors are normalized to get zero mean and unity standard deviation over the training set. The feedforward networks use log-sigmoid transfer functions for all hidden units in their hidden layer and linear transfer functions at the output layer. Biases and weights of each unit are initialized to very small random values.

4.2. Network learning algorithms

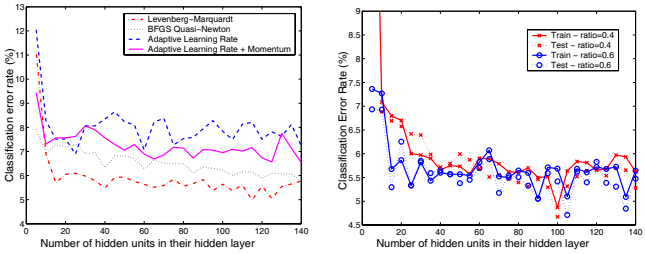
To avoid overfitting in the backpropagation learning algorithm, a weight decay heuristic (regularization) is used to decrease each weight by some small factor during each iteration. This modifies the typical performance function by adding a penalty term corresponding to the sum of squares of the network weights [9]:

$$Ereg = \gamma \frac{1}{N} \sum_{i=1}^{N} (t_i - o_i)^2 + (1-\gamma) \frac{1}{M} \sum_{j=1}^{M} (w_j)^2 \qquad (13)$$

where γ is a performance ratio, w_j are the weights of the NN, and t_i and o_i are the output and target values, respectively.
This approach results in smaller weights and biases and forces the network response to be smoother over its complex decision surface [8]. We select the best learning algorithm among the following ones: momentum, variable learning rate, Levenberg-Marquardt and BFGS Quasi-Newton algorithms. Some common parameters of the learning algorithms are set as follows:

- The learning rate $lr = 0.05$.
- The number of iterations $epochs = 1000$.
- The training performance $goal = 1e-5$
- The performance ratio and the number of hidden units are varied as $\gamma = [0.3, 0.35, ..., 0.8]$ and $nH = [5, 10, ..., 140]$ to find out the optimal values where the sum of training and testing error rate is smallest.

The datasets are taken from the TIMIT database, dialect speaking region 1 (DR1). Each dataset is divided into 70% training set and 30% test set. Female speaker and male speaker datasets are collected separately to investigate a gender-dependency of the phonetic classifier. Another mixed dataset containing both of genders is used to design a gender-independency phonetic classifier.



(a) the best learning algorithm at $\gamma = 0.5$

(b) the best performance ratios for LM algorithm

Fig. 3: The average error percentage on the training set and the test sets for the three-classes NN classifier and for female speaker.

From the results of the training phase, we see that the Levenberg - Marquardt (LM) algorithm gives the highest classification performance generally (Fig. 3-a). For this learning algorithm, the optimal choice of the performance ratio γ and number of the hidden units nH, which achieves the lowest error percentage on the training and test sets of the one-class NN classifier and three-classes NN classifier for female speaker is $0.75 - 60$ and $0.40 - 100$ (Fig. 3-b), for male speaker is $0.55 - 55$ and $0.50 - 125$, and for mixed-speaker is $0.55 - 70$ and $0.65 - 110$.

5. COMBINED CLASSIFICATION ALGORITHM

A phonetic group classification algorithm is proposed with four sequential steps shown in Fig. 4.

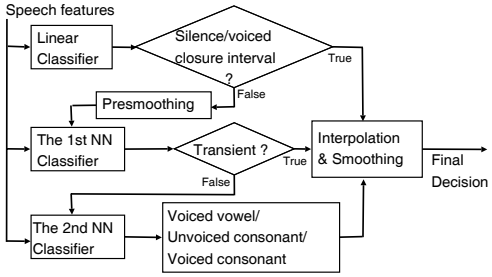


Fig. 4: The classification combined algorithm.

• First, silence (S) and voiced closure interval frames are detected by linear classifier (see Fig. 5) using threshold-based decision model in [1], with $AWER3 = 99\%$, $EVD1 = 0.15$, $EVD2 = 0.35$, $SAE1 = 0.001/160$, $SAE2 = 0.025/160$ and $SAE3 = 0.016/160$ (where SAE3 is a new threshold suggested to improve silence detection). Subsequently, presmoothing is used to eleminate some wrong decisions to decrease probability that the transient classifier can make incorrect decisions.

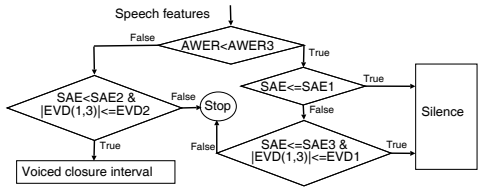


Fig. 5: The flow chart of the linear classifier in [1].

• Second, transient (T) frames are detected by considering frames which immediately follow silence or voiced closure interval frames. Five parameters such as $AWER$, ZCR, $CTDR(1)$, $CTDR(2)$, and $CTDR(3)$ are computed for these frames and classified by the first NN. The network distinguishes between transient frames and other frames.

• Third, four parameters such as $AWER$, SAE, ZCR and $EVD(1,3)$ of every not yet classified frame are calculated to build the input vectors for the second NN. The three classes voiced vowel (V), voiced consonant, and unvoiced consonant are recognized based on the output values of the second neural network.

• Finally, an interpolation method relying on phonemic features is implemented to build the temporal boundary of plosives and affricates which are formed by closure interval + transient + stop release frames which are the detected unvoiced consonant frames following transient frame. Then, some remaining incorrect decisions are repaired by the smoothing method based on sequential consistency of speech sound such as: VVSVV → VVVVV, or SSTSS → SSSSS, etc..

6. CLASSIFICATION RESULTS

Speech sounds used to build the experiment dataset are general data including transition frames and extracted from the DR1, TIMIT database with 7 male and 4 female speakers, 110 utterances, 85 sentences and 30065 frames in total. The performance of our new algorithm is assessed on the test set in the following ways:

• The relative classification error = (misclassified frames*100%) / tested frames [%].

The classification results are reported in Tab. 1 and Tab. 2. We observe a good generalization performance of the NN classifiers because the error rates are similar both for the training and test data, specially for unvoiced consonants.

	Silence			Voiced closure interval		
	M	F	A	M	F	A
Train	1.82	2.38	2.10	8.79	8.83	8.81
Test	1.95	2.52	2.23	9.94	9.56	9.75

Table 1: Classification error percentage corresponds with male speaker (M), female speaker (F) and average values (A).

		Gender-Dependent		Gender-Independent	
		Train	Test	Train	Test
Transient	M	5.91	6.29	6.45	6.92
	F	5.42	6.94	4.82	5.56
	A	5.67	6.93	5.63	6.55
Voiced vowel	M	2.30	2.67	2.50	3.02
	F	1.29	2.09	2.34	2.85
	A	1.79	2.38	2.42	2.93
Unvoiced consonant	M	1.59	1.69	2.36	2.48
	F	1.15	1.29	1.89	1.99
	A	1.37	1.49	2.12	2.23
Voiced consonant	M	13.77	14.98	15.26	16.36
	F	10.31	11.08	11.34	13.85
	A	12.04	13.03	13.30	15.11

Table 2: Classification error percentage of four classes with gender-dependent and gender-independent classifier.

The average difference of error percentages between two types of classifiers is 1.11%. It is lower than the one using hybrid features with NN (2%) in [5]. We see that the results achieved by the gender-independent NN are somehow worser than the results got by the gender-dependent NN. The reason is the gender-independent NN has to learn more complex borders of the classes from the bigger mixtured-dataset.

In comparison with our MTD model reported in [1], the new algorithm gets slightly better performance for the detection of silence, voiced closure interval, unvoiced consonants, and almost 2% lower error rate for transient and voiced consonants. We found that the error percentage of voiced vowels is slightly higher than in [1] but it is acceptable because we benefit more from the improved voiced consonants detection capabilities. In comparison with SUB-CRA in [2], it is clear that our new algorithm gains lower error rate for silence and unvoiced consonants (2.42% and 1.17% in comparison with 3.5% and 4.83%, respectively).

The average error percentage of 3 classes silence, voiced vowels and unvoiced consonants is 1.25% lower than the one of classifier using a NN with 5 features in [6]. In comparison with the classifier using high-rank function neural network in [4], our combined classifier get 0.95% higher average error rate for voiced vowels and unvoiced consonants. This is resonable because voiced consonants are not detected by [4]. That means our new classifier has wider classification ability.

• The error rate of plosives and affricates is calculated with speech datas in DR4, TIMIT (as used in [3]). Our error percentage approximates with the one using the optimal-filter based algorithm in [3] which detects only stop consonant (16.78% in comparison with 16%), and is 3.76% lower than the one in [1].

• The average error rate difference between male speaker and female speaker is reduced with 1.14% compared to 1.25% of the MTD in [1] and 1.33% of the EGG-based algorithm in [7]. This reduction is useful in approaching towards a gender-independent speech classification.

7. CONCLUSION AND OUTLOOK

The experimental results presented in the paper illustrate that, in general, the neural network classifiers obtained with an advanced training algorithm typically produce the better and more robust performance than other non-linear classifiers for clean speech as represented by the TIMIT database. The influence of environmental noise such as car noise, street noise and white noise will be considered carefully in our future research. This is usefull for speech classification in hard environments.

8. REFERENCES

[1] Pham Van Tuan, Gernot Kubin, "DWT-based Classification of Acoustic-Phonetic Classes and Phonetic Units", *Proc.ICSLP'04.*, Jeju, South Korea, Oct. 2004.

[2] Z. Lachiri, N. Ellouze, "Speech classification in noisy environment using subband decomposition", *Proc. ISSPA*, Vol. 1, pp. 409-412, 2003.

[3] P. Niyogi, M. M. Sondhi, "Detecting stop consonant in continuous speech", *J. Acoust. Soc. Am.*, Vol. 111, pp. 1063-1076, 2002.

[4] Jiang Minghu, Yuan Baozong, Lin Biquin, "The consonant/vowel speech classification using high-rank function neural network", *Proc. ICSP*, Vol.2, pp. 1469-1472, Brighton, UK, 1996.

[5] Yingyong Qi, Bobby R. Hunt, "Voiced-Unvoiced-Silence Classifications of Speech Using Hybrid Features and a Network Classifier", *IEEE Trans. on Speech and Audio Process.*, Vol. 1, pp. 250-255, 1993.

[6] Thea Ghiselli-Crippa, Amro El-Jaroudi, "Voiced-unvoiced-silence classification of speech using neural nets, *IJCNN*, Vol. 2, pp. 851-856, Seattle, USA, 1991.

[7] D.G. Childers, M. Hahn, J.N. Larar, "Silence and voiced/unvoiced/mixed excitation classification of speech", *IEEE Trans. on Acoust, Speech, Signal Process.*, Vol.37, No.11, pp. 1771-1774, 1989.

[8] Tom M. Mitchell, "Machine Learning", *McGraw-Hill*, USA, 1997.

[9] Matlab TM, "Toolbox of Neural Network Reference Guide", *The MathWorks Inc.*, 1995.