VOICED/UNVOICED SPEECH DISCRIMINATION IN NOISE USING GABOR ATOMIC DECOMPOSITION

Arthur P. Lobo and Philipos C. Loizou

Department of Electrical Engineering
University of Texas at Dallas
Richardson, TX 75083-0688, USA
Email: {arthur.lobo, loizou}@utdallas.edu

ABSTRACT

A new algorithm is developed for voiced-unvoiced speech discrimination in noise. Short segments of speech are modeled as a sum of basis functions from a Gabor dictionary. In each iteration, a Gabor atom is fitted (using the matching pursuit algorithm) to the residual obtained by subtracting the best-fit Gabor atom from the previous residual. Multiple discriminant analysis is used to reduce the dimensionality of the vector of Gabor coefficients to give a low-dimensional feature vector for classification. A Radial Basis function neural network is trained on the reduced feature vector set to discriminate between voiced and unvoiced speech/silence segments. On a database of 62 sentences in 5-dB SNR speech-shaped noise, 84% correct classification accuracy was obtained.

1. INTRODUCTION

The problem of voiced/unvoiced speech determination is an important one and has been worked on extensively by researchers [1]-[6] during the last three decades. In [1,2] a statistical parametric method was proposed whereas in [3,4,5] non-parametric methods based on linear discrimination functions, multi-layer feedforward and recurrent neural networks were adopted. In [6], a two-channel approach which made use of the speech and electroglottogram signals was pursued.

Most of the above methods proposed for voiced/unvoiced classification were implemented and tested in quiet. Voiced/unvoiced classification in noise, however, is a far more challenging task since the noise can potentially mask low-energy speech segments such as fricatives (e.g., /f/, /th/) and stop-consonants (e.g., /b/, /d/). Also, most of the above methods utilized long analysis frames with some [2] using as large as 40-ms duration frames.

In this paper, we propose an algorithm for voiced/unvoiced speech discrimination in noise that is based on the Gabor atomic decomposition of the speech waveform. The Gabor representation was chosen because it uses a family of functions that are well localized in both time and frequency. Such a representation received only a limited attention in speech processing. Gabor atomic decomposition for audio signal enhancement has previously been suggested by Wolfe *et al.* [7] and has been applied to radar target recognition by Shi and Zhang [8]. The well-localized, in time and frequency, properties of the Gabor functions allow us to analyze the speech signal using short-duration segments. In this paper, we investigated voiced/unvoiced classification of 3.2-ms duration frames.

This paper is organized as follows. Section 2 describes the proposed algorithm, Section 3 presents the experimental results and Section 4 presents our conclusions.

2. PROPOSED VOICED/UNVOICED CLASSIFICATION ALGORITHM

The speech signal $f(t)$ can be represented as:

$$\hat{f}(t) = \sum_{n=1}^{M} a_n g_{\gamma n}(t) \qquad (1)$$

where M is the order of decomposition, a_n are the Gabor coefficients and $g_{\gamma n}(t)$ are the basis functions, also called the Gabor atoms. A Gabor atom consists of a cosine-modulated Gaussian window function:

$$g_\gamma(t) = \frac{1}{\sqrt{s}} g\left(\frac{t-u}{s}\right)\cos(vt+w) \qquad (2)$$

where $g(t) = e^{-\pi t^2}$ is the Gaussian window function and $\gamma = (s, u, v, w)$ are the time-frequency parameters. The function $g_\gamma(t)$ is centered at u and its energy is mostly concentrated in a neighborhood of u whose size is proportional to s. The space

ICASSP 2003

of time-frequency parameters can be discretized as $\gamma = \left(a^j, pa^j \Delta u, ka^{-j} \Delta v, i \Delta w \right)$, with $a = 2$, $\Delta u = 1/2$, $\Delta v = \pi$, $\Delta w = \pi/6$, $0 < j \leq \log_2 N$, $0 \leq p < N2^{-j+1}$, $0 \leq k < 2^{j+1}, 0 \leq i \leq 12$ to form the so called Gabor dictionary. Here N is the number of samples in a frame.

As discussed in [7], the Gabor dictionary is highly redundant, and a regression model was used in [7] to reduce the Gabor dictionary. In this paper, we use the matching pursuit algorithm, proposed in [9], to prune the Gabor dictionary. The matching pursuit algorithm is a greedy algorithm that chooses at each iteration a waveform that is best adapted to an approximate part of the signal and hence is locally adaptive. It has been shown [9] that for highly non-stationary signals (e.g., plosives), the matching pursuit algorithm performs better than a globally adaptive algorithm like that proposed by Coifman and Wickerhauser [10], which selects the basis best adapted to the global signal properties.

If $R^0 f = f$ denotes the signal being modeled, the residual at the n-th iteration, denoted by $R^n f$, is given by:

$$R^n f = a_n g_{\gamma n} + R^{n+1} f \qquad (3)$$

where a_n are the Gabor coefficients [11,12] computed as the inner product $a_n = \left\langle R^n f, g_{\gamma n} \right\rangle$. The norm of the residual ($\left\| R^{n+1} f \right\|$) assesses the degree of fitness of the Gabor atom $g_{\gamma n}$ to the residual at iteration n. In each iteration, one Gabor atom is added to the approximation to model the residual. The procedure continues iteratively until a prescribed number, M, of Gabor expansion coefficients are generated. The dictionary element chosen at each stage is the element that provides the greatest reduction in mean square error between the true signal $f(t)$ and the approximated signal $\hat{f}(t)$. In this sense, the speech signal structures are coded in order of importance. This property is desirable also in cases where the bit budget is limited, such as in low rate speech coding.

The resulting M Gabor coefficients are used in voiced/unvoiced classification. Multiple discriminant analysis [13] is first used to reduce the dimensionality of the vector of Gabor coefficients. The vector of Gabor coefficients x, of dimension M, is projected to a $(c-1)$ dimensional space y_i where c is the number of classes. The projection is done as:

$$y_i = w_i^T x, \quad i = 1,....,c - 1 \qquad (4)$$

where w_i are the eigenvectors which satisfy the generalized eigenvalue problem:

$$S_B w_i = \lambda_i S_w w_i \qquad (5)$$

where S_B is the *between-class* scatter matrix, S_w is the *within-class* scatter matrix and λ_i is the i-th eigenvalue.

Finally, the reduced dimensional feature vector y_i is input to a Radial Basis Function Neural Network (RBFNN) [14,15,16] with Gaussian units that are trained to map the feature space to the output (class) space. The RBFNN was used because the weight matrix between the Gaussian units and the output units can be estimated using all the training patterns at one time. This is in contrast to the back-propagation algorithm which uses an incremental training approach to train multi-layer perceptrons.

3. EXPERIMENTAL RESULTS

The performance of the proposed voiced/unvoiced classification algorithm was evaluated using 62 sentences uttered by a male speaker. The sentences (sampled at 20 kHz) were taken from the HINT database [17]. Speech-shaped noise [17] at 5-dB SNR was added to the clean speech waveforms to create the noisy speech waveforms. The data was manually segmented into voiced and unvoiced segments with a frame size of 64 samples (3.2-ms). A vector of M=40 Gabor coefficients was computed for each frame. The vector was projected to a c-1 dimensional space using multiple discriminant analysis. Here c=2, where the first class corresponds to vowels, nasals and glides and the second class corresponds to plosives, fricatives and silence. The reduced feature vectors (features were scalars since c=2) were divided into training and test sets. The training set consisted of 29,206 patterns and the test set consisted of 3,962 patterns. A RBFNN with 15 hidden units and 2 output units corresponding to the voiced and unvoiced classes, was trained on the training set and tested on the test set.

The centers and the widths of the Gaussian units were initialized by dividing the training set into 15 consecutive segment clusters and finding the feature mean and standard deviation for each cluster. The outputs of the RBFNN were low pass filtered with a cutoff frequency of 15 Hz before the decision step. A voiced determination was made when the output of the voiced RBFNN unit was greater than that of the unvoiced unit, otherwise an unvoiced/silence determination was made.

When tested on 7 sentences (3,962 patterns), the RFFNN classification accuracy was 84%. Figure 1 shows the percent correct classification as a function of the size of the training set for a 15-unit RBFNN. As can be seen, as the number of training patterns increased, the
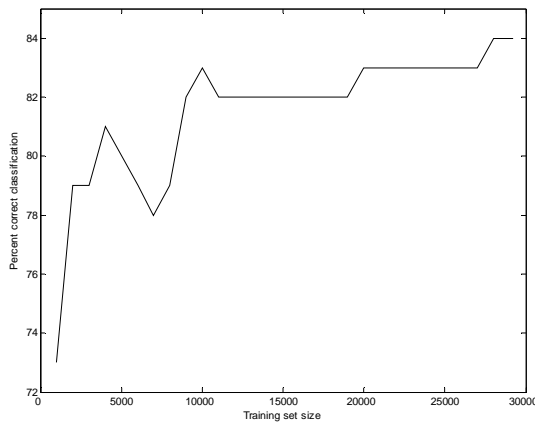
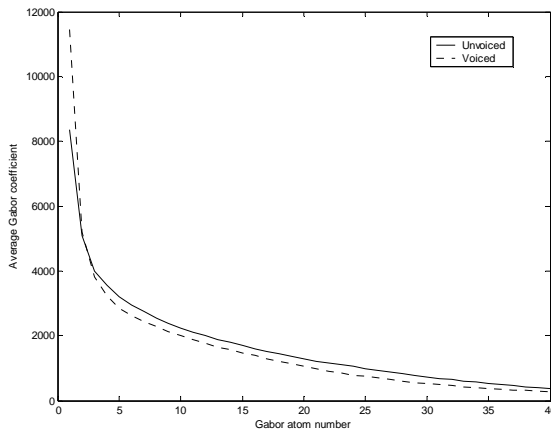Figure 1. RBFNN percent correct classification as a function of the training set size.

Figure 2. Plot of absolute Gabor coefficients vs Gabor atom number.

percent correct classification increased and saturated at 84%.

Figure 2 shows the absolute value of the Gabor coefficient plotted as a function of the Gabor atom number for voiced and unvoiced speech/silence frames. The Gabor coefficients were averaged over 570 frames of a sentence file. As can be seen, the Gabor coefficients for unvoiced speech/silence frames are smaller in value than the coefficients of voiced speech frames for Gabor atom numbers less than 3. For Gabor atoms numbers greater than 3, the coefficients for the unvoiced speech frames are greater than the coefficients for the voiced speech frames. This separation of the two classes in feature space is implemented by the modeling of the data clusters by the RBFNN.

Figure 3 shows an example output of the RBFNN trained on the Gabor coefficients and the true voiced/unvoiced decisions for the test sentence "They watched a scary movie". Figures 3(a) and 3(b) show the
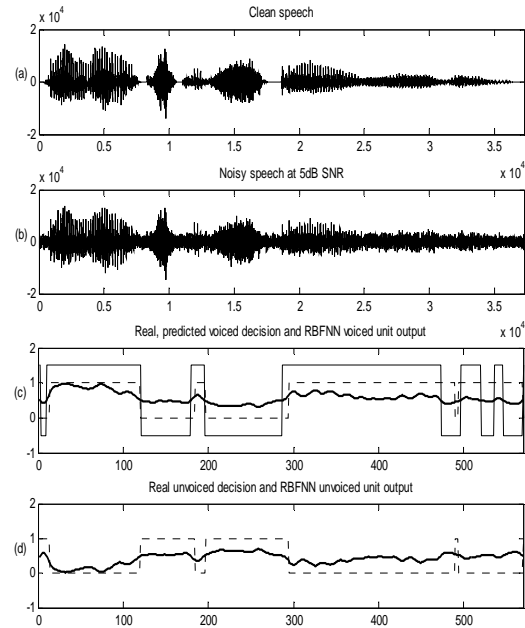

Figure 3. Time-amplitude waveforms of (a) the clean speech signal, (b) the noisy speech signal, (c) true voiced decision (dashed line), predicted voiced decision (thin solid line) and RBFNN voiced unit output (thick solid line), (d) true unvoiced decision (dashed line) and RBFNN unvoiced unit output (thick solid line).

clean and noisy speech waveforms respectively with sample number on the abscissa and amplitude on the ordinate. Figure 3(c) shows the true (dashed line) and predicted (thin solid line) voiced decision and RBFNN voiced unit output (thick solid line). Figure 3(d) shows the true unvoiced decision (dashed line) and RBFNN unvoiced unit output (thick solid line). A hard decision value greater than 0 corresponds to a true decision while a hard decision value less than 0 corresponds to a false decision.

4.   SUMMARY AND CONCLUSIONS

A new method for voiced/unvoiced speech discrimination in noise was developed. The proposed algorithm was based on the matching pursuit algorithm that generates a vector of Gabor coefficients. Multiple discriminant analysis is applied to this set of coefficients to get a reduced dimensional feature vector for classification. A Radial Basis function neural network is trained on the set of reduced feature vectors as a voiced/unvoiced speech classifier. High classification accuracy using a small frame size (3.2-ms) was obtained for speech sentences embedded in 5-dB SNR speech-shaped noise.

6. REFERENCES

[1] B. Atal, and L. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition," *IEEE Trans. on ASSP*, vol. ASSP-24, pp. 201-212, 1976.

[2] S. Ahmadi, and A.S. Spanias, "Cepstrum-Based Pitch Detection using a New Statistical V/UV Classification Algorithm," *IEEE Trans. Speech Audio Processing*, vol. 7 No. 3, pp. 333-338, 1999.

[3] Y. Qi, and B.R. Hunt, "Voiced-Unvoiced-Silence Classifications of Speech using Hybrid Features and a Network Classifier," *IEEE Trans. Speech Audio Processing*, vol. 1 No. 2, pp. 250-255, 1993.

[4] L. Siegel, "A Procedure for using Pattern Classification Techniques to obtain a Voiced/Unvoiced Classifier," *IEEE Trans. on ASSP*, vol. ASSP-27, pp. 83-88, 1979.

[5] T.L. Burrows, *Speech Processing with Linear and Neural Network Models*, Ph.D. thesis, Cambridge University Engineering Department, U.K., 1996.

[6] D.G. Childers, M. Hahn, and J.N. Larar, "Silent and Voiced/Unvoiced/Mixed Excitation (Four-Way) Classification of Speech," *IEEE Trans. on ASSP*, vol. 37 No. 11, pp. 1771-1774, 1989.

[7] P. J. Wolfe, S.J. Godsill, and M. Dörfler, "Multi-Gabor Dictionaries for Audio Time-Frequency Analysis," *Proc. IEEE Wkshp. on Appl. of Sig. Proc. to Audio and Acoust.*, New Paltz, NY, pp. 43-46, 2001.

[8] Y. Shi, and X.D. Zhang, "A Gabor Atom Network for Signal Classification with Application in Radar Target Recognition," *IEEE Trans. on Signal Processing*, vol. 49 No. 12, pp. 2994-3004, 2001.

[9] S.G. Mallat, and Z. Zhang, "Matching Pursuit with Time-Frequency Dictionaries," *IEEE Trans. on Signal Processing*, vol. 41 No. 12, pp. 3397-3415, 1993.

[10] R.R. Coifman, and M.V. Wickerhauser, "Entropy-based Algorithms for Best Basis Selection," *IEEE Trans. Informat. Theory*, vol. 38, pp. 713-719, 1992.

[11] M.J. Bastiaans, "Gabor's Expansion of a Signal into Gaussian Elementary Signals," *Proc. of the IEEE*, vol. 68, pp. 594-598, 1980.

[12] S. Qian, and D. Chen, "Signal Representation using Adaptive Normalized Gaussian Functions," *Signal Processing*, vol. 36, pp. 1-11, 1994.

[13] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, 2nd ed., John Wiley, 2001.

[14] T. Poggio, and F. Girosi, "Networks for Approximation and Learning," *Proc. of the IEEE*, vol. 78, pp. 1481-1497, 1990.

[15] D.S. Broomhead, and D. Lowe, "Multivariable Functional Interpolation and Adaptive Networks," *Complex Systems*, 2, pp. 321-355, 1988.

[16] S.S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, 1998.

[17] M. Nilsson, S.D. Soli, and J.A. Sullivan, "Development of the Hearing in Noise Test for the Measurement of Speech Reception Thresholds in Quiet and in Noise," *J. Acoust. Soc. Am.*, 95, pp. 1085-1099, 1994.