

Wavelet Formants Speaker Identification Based System via Neural Network

K. Daqrouq¹, Emad Khalaf¹, A. Al-Qawasmi¹, and T. Abu Hilal²

¹Philadelphia University/Communications and Electronics Dept., Amman, Jordan
haleddaq@yahoo.com, emad_khalaf1@yahoo.com, qawasmi@philadelphia.edu.jo,

²Dhofar University, Oman
tr_hilal@yahoo.com

Abstract— In this paper Discrete wavelet Transform with logarithmic Power Spectrum Density (PSD) are combined for speaker formants extraction, to be used as evident classification features. For classification, Feed Forward Back Propagation Neural Network FFBNN method is proposed. The Discrete Wavelet formants Neural Network DWFNNT system works with excellent capability of features tracking even with 0dB SNR. Text - dependant system is used, so that the system can be applied in password or PINs identification in any security system. The proposed system is compared with K-means algorithm based clustering method. The results show excellent performance with 93.21% Recognition Rate (RR).

Index Terms— wavelet, speech signal, formants, neural network, speaker identification.

I. INTRODUCTION

Over last four decades many solutions of speaker recognition have been appeared in literatures [1,11-14]. The Al-Alaoui algorithm for pattern classification [1,5,6,7] was motivated by Patterson and Womack's [9] and Wee's [10] proofs that the Mean Square Error (MSE) solution of the pattern classification solution gives a minimum mean-square-error approximation to Bayes' discrimination, weighted by the probability density function of the sample. All audio techniques start by converting the raw speech signal into a sequence of acoustic feature vectors carrying distinct information about the signal. This feature extraction is also called "front-end" in the literature. The most commonly used acoustic vectors are Mel Frequency Cepstral Coefficients (MFCC) [15, 16], Linear Prediction Cepstral Coefficients (LPCC) [17-19], and Perceptual Linear Prediction Cepstral (PLPC) Coefficients. Text-dependent methods are usually based on template-matching techniques. In this approach, the input utterance is represented by a sequence of feature vectors, generally short-term spectral feature vectors. The time axes of the input utterance and each reference template or reference model of the registered speakers are aligned using a dynamic time warping (DTW) algorithm and the degree of similarity between them, accumulated from the beginning to the end of the utterance, is calculated [2,3,4,8].

A method using statistical dynamic features has recently been proposed. In this method, a multivariate auto-regression (MAR) model is applied to the time series of cepstral vectors and used to characterize speakers [20,21,28].

K-means is popular clustering algorithm that has been used in a variety of application disciplines, such as image clustering [33] and information retrieval [34], as well as, speech and speaker identification. In [34], modified version for background knowledge was of significant use to the clustering community. A genetic algorithm-based efficient clustering technique that utilizes the principles of K-Means algorithm is described in [35]. In [37], the syllable contour is classified into several linear loci that serve as candidates for the tone-nucleus using segmental K-means segmentation algorithm. In [38], the problem of slow speaker identification for large population systems is considered by using of K-means clustering algorithm.

In this research, the PSD and Discrete Wavelet Transform based recognition system DWFNNT is proposed. The proposed system is divided into two core steps: 1) Features Extracting by PSD and Wavelet Transform and 2) Classification using FFBNN. Paper contains 7 parts: Introduction, Proposed Method, Formants Extraction by Power Spectrum Density, WP_{XX} for Speaker Classification, Feed Forward Back Propagation Neural Network, Results and Discussion, and finally Conclusion.

II. PROPOSED METHOD

In this paper wavelet transform based speaker identification system is presented. The system based on two main stages: 1.Feature Extraction and 2.Speaker Classification. In the first stage, speech signal is decomposed into Discrete Wavelet Transform (DWT) Approximation Coefficients via J levels (J sub-signals). All J DWT Approximation sub-signals are given to PSD, which is estimated using the Yule-Walker autoregressive (AR) method. The second stage contains Speaker Features Classification using FFBPNN. The impostor feature data is given to network to be trained by 1000 to 5000 epochs with four binary code target for each column input feature data. Then, recognition rate is calculated for each J sub-signals separately, by simulating the network with each model data stored in system memory to be verified. Afterwards, decision is taken. The Recognition Rate (RR) is calculated as the ratio of number of zeros (NZ) in network simulating results minus target with respect to the number of target elements N.

$$RR = \frac{\sum NZ(\text{sim}(\text{net}, \text{model data}) - \text{Target})}{N} \quad (1)$$

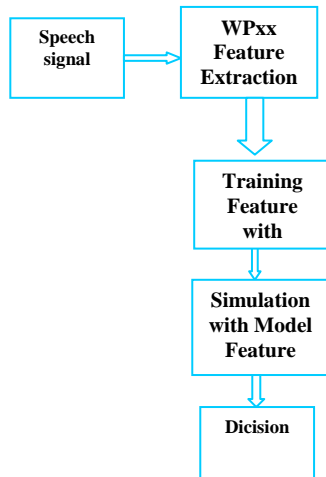


Figure 1. Block Diagram of proposed system

III. FORMANTS EXTRACTION BY POWER SPECTRUM DENSITY

Formants are the frequency parts of speech signal that are related to the human distinct vocal tract anatomy form, which is distinguishable for each person.

In this paper, we use these formants as the basic speaker features carriers [30], which is determined by PSD (is denoted as P_{XX}) and shown at Fig.2, which is estimated using the Yule-Walker autoregressive (AR) method. This method, also called windowed method, fits an AR linear prediction filter model to the signal by minimizing the forward prediction error in the least squares sense. This formulation leads to the Yule-Walker equations, which are solved by the Levinson-Durbin recursion [8, 9]. The spectral estimate returned by method is the squared magnitude of the frequency response of this AR model. Then N vector of the speaker's formants is represented by logarithmic scale:

$$F_{P_{XX}}(n) = \sum_n^N 10 \log_{10}(P_{XX}) \quad (2)$$

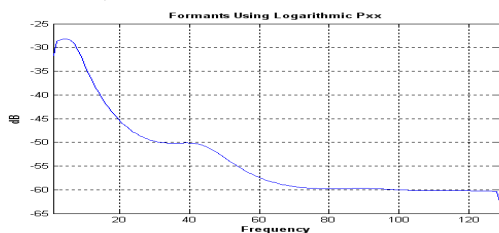


Figure 2. $F_{P_{XX}}$ speaker's formants represented by logarithmic scale

$F_{P_{XX}}(n)$ returns P_{XX} containing adequate features to exhibit the speaker uniqueness. To spread out the PSD competence of formants illustration, as well as speaker features extraction, we suggest the Discrete Wavelet

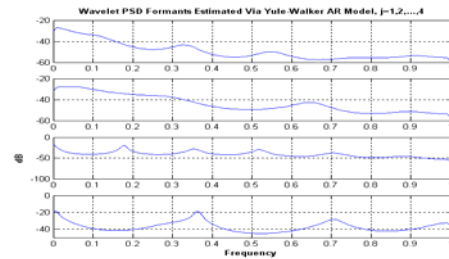


Figure 3. $WP_{XX}(j=1,2,\dots,4)$ of one utterance for same person

Transform Approximation Coefficients a_j of multiple scales as an input to P_{XX} , at that time P_{XX} output is denoted by WP_{XX} (Fig.3). Detail Coefficients d_j is mistreated.

$$a_{j+1}(t) = \sum_m h(m-2t)a_j(m) \quad (3)$$

Where, the set of numbers $a_j(m)$ represents the down sampled approximation of the signal at the resolution 2^{-j} (Fig.4). $h(n)$ is the coefficient of the linear combination that approximates the wavelet scaled version function $\phi(x)$ [22,23,24]:

$$\phi(x/2) = 2^{1/2} \sum_m h(m)\phi(x-m) \quad (4)$$

$$h(m) = \frac{1}{2^{1/2}} \int \phi(x/2)\phi(x-m)dx \quad (5)$$

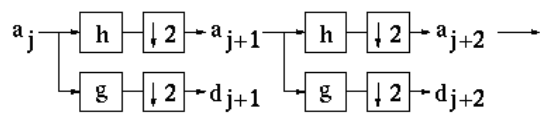


Figure 4. DWT coefficients generation

WP_{XX} assists significantly in formant re-demonstration in sharp form (Fig.4) as well as; expand the signal analysis form short-term to long-term, over all signal frequencies band-passes due to using multistage J.

IV. WPXX FOR SPEAKER CLASSIFICATION

Speech signal conveys linguistic message, and also hidden information about the speaker. These information that is called feature, is the unique speaker message contained in the spoken acoustic wave. Consider the number of non zeros samples over equal frames with respect to the whole signal non zeros samples number is Signal occurrences probability:

$$P_{occur} = \frac{NF}{NW} \quad (6)$$

Where, NF is the number of non zeros samples in one frame and NW is the number of non zeros samples in the whole speech signal.

Such probability can have unseen information about speaker. This is due to demonstrating attribute about

signal concentration over windows. Better presentation of such manner can be accomplished by WP_{XX} , where the signal is decomposed into orthogonal J levels of DWT sub signals via different band pass of frequency, similar to Occurrences Probability feature capturing approach, but in more flexible and reliable methodology. In this section, we analyse the effect of WP_{XX} at classification of the speech signal via J levels of DWT. Two quantity methods to determine the classification degree are utilized: Correlation coefficient (ρ), and Model-to- Impostor Ratio (MIR).

1. ρ is the expectation $E[.]$ of the product of the speech signal model X about mean value and speech signal impostor Y about mean value related to the product of the Standard Deviation of X (σ_X) and Standard Deviation of Y (σ_Y):

$$\rho = \frac{E[(X - \bar{X})(Y - \bar{Y})]}{\sigma_X \sigma_Y} \quad (7)$$

ρ is efficient likeness or similarity tool judgment of random Variables X and Y in term of unity (for one hundred percent similarity). Fig. 5 shows the results of ρ values that were calculated for WP_{XX} of speaker model signal and four speaker signals, three of them belong to the same speaker and fourth one is impostor. The results were calculated over $j=1,2,\dots,6$. ρ is the smallest for impostor signal.

2. MIR is proposed in this paper as a dB measure of speech model signal and impostor ratio:

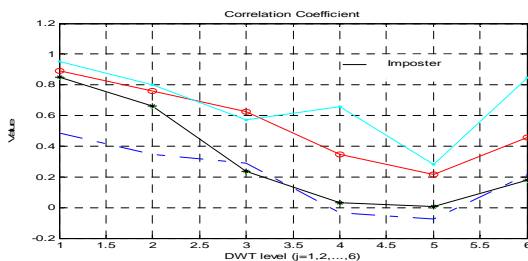


Figure 5. ρ values that were calculated for WP_{XX} of impostor speaker signal and three same speaker signals

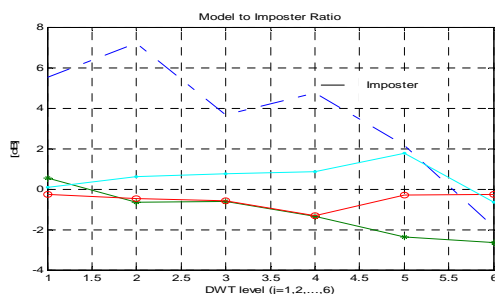


Figure 6. MIR values calculated for WP_{XX} of impostor speaker signal and three same speaker signals

$$MIR = 20 \log_{10} \frac{\sum_{-\infty}^{\infty} X}{\sum_{-\infty}^{\infty} Y} \quad dB \quad (8)$$

MIR calculates dB measure of above ratio. Now, if Y belongs to the same speaker the result should be near zero dB, otherwise [31], is away from zero, in term of absolute value (Fig.6). The results illustrated here show high classification capability using WP_{XX} . Fig.7 presents the results taken for the same signals used in Fig.5 and 6 contaminated with random noise (SNR= 0dB). The results confirm the possibility of discrimination even when tough noise has been added. Based on our experimental observations, we have noticed bad and confusing classification results by WP_{XX} of $j=6$ (at level 6).

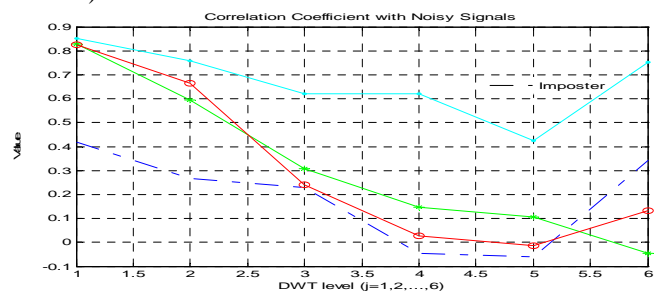


Figure 7. ρ calculated for WP_{XX} of noisy impostor signal and three noisy same speaker signals

Level 3 is weak but not confusing. We decided to utilize WP_{XX} of five levels only.

V. FEED FORWARD BACK PROPAGATION NEURAL NETWORK

Back-propagation is the most frequently used method for training multi-layer feed-forward networks for most networks. The learning process is based on a appropriate error function, which is then minimized with respect to the weights and bias. Hence, we can assess the derivative of the error with respect to weights, as well as these derivatives can then be used to come across the weights that minimize the error function, by either using the popular gradient descent or other optimization methods. The used algorithm for evaluating the derivative of the error function is known as back-propagation, because it propagates the errors backward through the network. From Matlab environment, we take *Newff* function, which constructs a feed forward back propagation network (Fig.8).

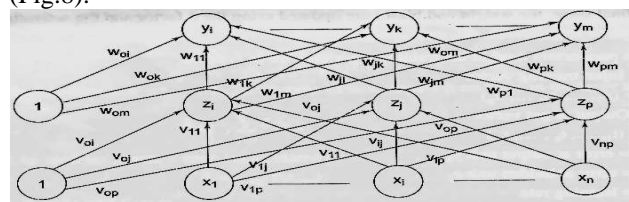


Figure 8. Architecture of back propagation network

VI. RESULTS AND DISCUSSION

Tested speech signals were recorded via PC-sound card, with spectral frequency 4000 Hz and sampling frequency 16000 Hz, over about 2 sec. time duration. Each speaker recorded Arabic expression "besme allah Alrahman Alraheem" that means in English "In the Name of God" that was recorded one time by the speaker. The speaker recorded 26 utterances. 4 females and 18 males got a part in utterances recording. From above speech signal description, we can notice that presented recognition system is text-dependent system [27], because prompts are common across all speakers that can share secrets (passwords or PINs). In order to create multi-factor authentication scenario, the speaker in each trail is compared to all models stored in database.

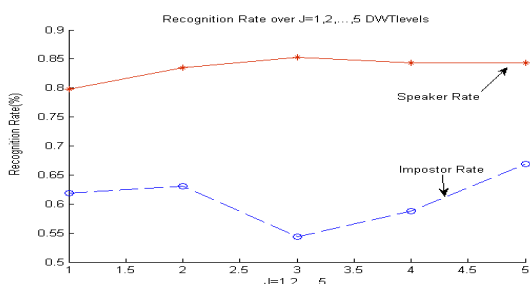


Figure 9. RR results of proposed system for speaker and impostor speaker signals

Fig. 9 presents the system Recognition Rate calculated for impostor input verified with model signal stored in system memory, as will as input signal verified with model signal of same person stored in system memory. We can notice the. We can notice the big gap between two results.

Table 1 presents RR results of four identification method of different classification concepts (CC) calculated for 480 numbers of testing signals (NTS). Firstly, is the proposed method DWFNNT. secondly, is the Continuous Wavelet formants Neural Network CWFNNT, which is based on using Continuous Wavelet Transform instead of DWT. Thirdly, is formants Neural Network method FNNT and finally is Discrete Wavelet formants K-means WFDKM. Where the Square Euclidean distances from each point to every centroid in the K clusters vector are calculated over WPXX. The results show that proposed method is nearly like WFDKM, but is superior based on fact that no neighborhood (NH) results have been observed unlike in case of WFDKM where is about 19%.

TABLE I.
RECOGNITION RATE (RR) RESULTS OF FOUR METHODS WITH RECOGNITION CONCEPTS (CC)

Method	CC	j	NTS	NH	RR
DWFNNT	FFBNN	1,2,...,5	480	0	93.21
CWFNNT	FFBNN	1	480	0	87.23
FNNT	FFBNN	0	480	0	89.40%
WFDKM	k-means	1,2,...,5	480	18.7%	93.7%

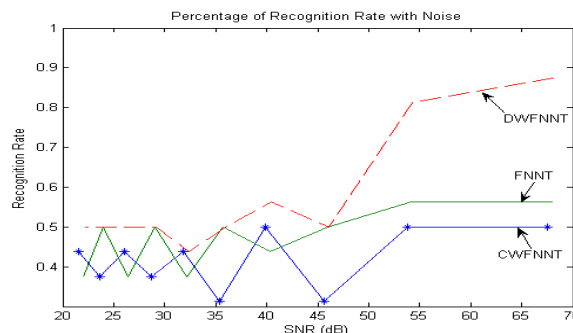


Figure 10. RR results of proposed system DWFNNT, CWFNNT and FNNT over different levels of noise

Fig. 10 presents RR results of proposed system DWFNNT, CWFNNT and FNNT over different levels of noise. This experiment is presented to illustrate a prove of system superiority in term of noise robustness. Finally, table 2 presents a comparison between DWFNNT and CWFNNT over j=1,2,3,...,10.

VII. CONCLUSIONS

In this Paper, Wavelet Transform based speaker feature extraction method is investigated by NNT. The introduced system in this paper depends on two steps features extraction by WPXX over five approximation DWT levels, due to its better capability of formants illustration over different band-pass of signal frequency. And classification based on FFBNN. The system works with excellent capability of features tracking even with 0dB SNR. Different concepts of comparison to other methods are proposed. Text - dependant system is used, so that the system can be applied in password or PINs identification in any security system, Banks, Hotel rooms, or other companies. More than four thousand trails were done to evaluate the system performance. The results show excellent performance about 94% classification rates.

REFERENCES

- [1] M. A. Al-Alaoui, Some Applications of Generalized Inverse to Pattern Recognition, Ph.D. Thesis, Electrical Engineering Department, Georgia Institute of Technology, December, 1974.
- [2] Elisabeth Zetterholm, Voice Imitation. A Phonetic Study of Perceptual Illusions and Acoustic Success. PhD thesis, Lund University, (2003).
- [3] <http://cslu.cse.ogi.edu/HLTsurvey/ch1node66.html#icassp89>, Maintained by Mike Noel and Wei Wei.
- [4] <http://www.itl.nist.gov/div893/biometrics/Biometris> from the movies. pdf, National Institute of Standards and Technology, Biometric Authentication Technology: From the Movies to Your Desktop by Fernando L. Podio1 and Jeffrey S. Dunn2,
- [5] M. A. Al-Alaoui, A New Weighted Generalized Inverse Algorithm for Pattern Recognition, IEEE Transactions on Computers, Vol. C-26, No. 10, pp. 1009-1017, October 1977.
- [6] M. A. Al-Alaoui, Application of Constrained Generalized Inverse to Pattern Recognition, Pattern Recognition, Pergamon Press, Vol. 8, pp. 277-281, 1976.

[7] M. A. Al-Alaoui, J. El Achkar, M. Hijazi, T. Zeineddine and M. Khuri, Application of Artificial Neural Networks to QRS Detection and LVH Diagnosis; Proceedings of ICECS'95, pp. 377-380, Amman-Jordan, 17-21 December 1995.

[8] Marple, S.L., Digital Spectral Analysis, Prentice-Hall, 1987, Chapter 7.

[9] Stoica, P., and R.L. Moses, Introduction to Spectral Analysis, Prentice-Hall, 1997.

[10] J. D. Ptterson and B. F. Womack, An Adaptive Pattern Classification System, IEEE Transactions Syst. Man. Cybern. , Vol. SSC-2, pp. 62-67, August 1966.

[11] W. G. Wee, "Generalized Inverse Approach to Adaptive Multiclass Pattern Classification, IEEE Transactions on Computers, Vol. C-17, pp.1157-1164,December 1968.

[12] Special Issue on Speaker Recognition, Digital Signal Processing, vol. 10, January 2000.

[13] R. Teunen, B. Shahshahani, and L. Heck, A Model-based Transformational Approach to Robust Speaker Recognition, ICSLP October 2000.

[14] A. Higgins, L. Bahler, and J. Porter, "Speaker Verification using Randomized Phrase Prompting," Digital Signal Processing, vol. 1, pp. 89-106,1991

[15] Chakroborty, S., Roy, A. and Saha, G., "Improved Closed set Text-Independent Speaker Identification by Combining MFCC with vidence from Flipped Filter Banks". International Journal of Signal Processing, Vol. 4, No. 2, Page(s):114-122, 2007.

[16] Sandipan Chakroborty* and Goutam Saha, Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter. International Journal of Signal Processing 5;1 © www.waset.org Winter 2009.

[17] Sanderson S. Automatic Person Verification Using Speech and Face Information. PhD thesis. Griffith University. 2002.

[18] Petry A. and Barone D. A. C. Fractal Dimension Applied to Speaker Identification. ICASSP (Salt Lake City). May 7-11. 405-408, 2001.

[19] Liu C. H., Chen O. T. C. A Text-Independent Speaker Identification System Using PARCOR and AR Model. MWSCAS. Vol 3, 332-335. 2002.

[20] D. Gabor, Theory of communication, Journal of I.E.E. 93 pp 429-441, 1946.

[21]

[22] P. Goupillaud, A. Grossmann, J. Morlet, "Cycle-octave and related transforms in seismic signal analysis", Geoexploration, 23, 85-102, 1984-1985.

[23] A. Grossmann and J. Morlet, Decomposition of Hardy functions into square integrable wavelets of constant shape, SIAM J. Math. Anal, Vol. 15, pp 723-736, 1984.

[24] Meyer, Wavelets, Ed. J.M. Combes et al., Springer Verlag, Berlin, p. 21, 1989.

[25] Abdel-Rahman Al-Qawasmi and Khaled Daqrouq "Discrete Wavelet Transform with Enhancement Filter for ECG Signal", WSEAS transaction on Biology and Biomedicine" 2009

[26] <http://www.dtic.upf.edu/~xserra/cursos/TDP/referencias/Park-LPC-tutorial.pdf>

[27] T. Matsui and S. Furui. Concatenated phoneme models for text-variable speaker recognition. IEEE Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing, Minneapolis, Minnesota, April 1993, pages 391--394.

[28] T. Matsui and S. Furui. Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. In IEEE Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing, Minneapolis, Minnesota, April 1993,], pages 157--160.

[29] C. Griffin, T. Matsui, and S. Furui. Distance measures for text-independent speaker recognition based on MAR model. IEEE Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing, Adelaide, Australia, April 1994, pages 309--312.

[30] J. M. Naik, L. P. Netsch, and G. R. Doddington. Speaker verification over long distance telephone lines. In IEEE Proceedings of the 1989

[31] International Conference on Acoustics, Speech, and Signal Processing, Glasgow, Scotland, May 1989, pages 524--527.

[32] X. Xiao, Normalization of the Speech Modulation Spectra for Robust Speech Recognition, IEEE Transaction on Audio, Speech, and Language Processing, Vol. 16, NO. 8, November 2008.

[33] Y.Chao,W.Tsai, H.Wang and R.Chang. Improving the characterization of the Aleternative hypothesis via minimum verification error with application to speaker verification, Pattern Recognition 42 (2009) 1351-1360.

[34] Marroquin, J., & Girosi, F. Some extensions of the k-means algorithm for image segmentation and pattern recognitionAI Memo 1390). Massachusetts Institute of Technology, Cambridge, MA.1993.

[35] Bellot, P., & El-Beze, M. A clusterin method for information retrieval (Technical Report IR-0199). Laboratoire d'Informatique d'Avignon. France. 1999.

[36] Kiri Wagsta, Claire Cardie, Constrained K- means Clustering with Background Knowledge, Proceedings of the Eighteenth International Conference on Machine Learning, p. 577-584, 2001.

[37] S.Bandyopadhyay,U.Maulik, An evolutionary technique based on K-Means algorithm for optimal clustering in RN, Information Sciences 146 221–237, 2002.

[38] J.E. Rougui, M. Gelgon, D. Aboutajdine, N. Mouaddib, M. Rziza, Organizing Gaussian mixture models into a tree for scaling up speaker retrieval. Pattern Recognition Letters 28 ,1314–1319, 2007.

[39] J. Zhang, K. Hirose, Tone nucleus modeling for Chinese lexical tone recognition, Speech Communication 42, 447–466, 2004.

[40] V.R.Apsingekar and P.L.De Leon, Speaker Model Clustering for Efficient Speaker Identification in Large Population, IEEE Transaction on Audio, Speech and Language Processing, Volume.

[41] <http://www.dtic.upf.edu/~xserra/cursos/TDP/referencias/Park-LPC-tutorial.pdf>

TABLE II.
RECOGNITION RATE (RR) RESULTS OF DWFNNT AND CWFNNT OVER J=1,2,3,...,10

Method	j=1	j=2	j=3	j=4	j=5	j=6	j=7	j=8	j=9	j=10	Avr.
DWFNNT	100	100	81.25	100	100	93.75	100	100	87.50	100	96.25
CWFNNT	100	93.75	93.75	93.75	100	93.75	81.25	87.50	93.75	93.75	93.13