Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 61 (2015) 416 – 421

Procedia Computer Science

Complex Adaptive Systems, Publication 5
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2015-San Jose, CA

A Wavelet Packet and Mel-Frequency Cepstral Coefficients-Based Feature Extraction Method for Speaker Identification

Claude Turner[a], Anthony Joseph[b]*

[a]*Dept. of Computer Science, Norfolk State University, 700 Park Ave, Norfolk, VA 23504*
[b]*Department of Computer Science, Pace University, 163 William St., New York, NY 10038*

Abstract

One of the most widely used approaches for feature extraction in speaker recognition is the filter bank-based Mel Frequency Cepstral Coefficients (MFCC) approach. The main goal of feature extraction in this context is to extract features from raw speech that captures the unique characteristics of a particular individual. During the feature extraction process, the discrete Fourier transform (DFT) is typically employed to compute the spectrum of the speech waveform. However, over the past few years, the discrete wavelet transform (DWT) has gained remarkable attention, and has been favored over the DFT in a wide variety of applications. The wavelet packet transform (WPT) is an extension of the DWT that adds more flexibility to the decomposition process. This work is a study of the impact on performance, with respect to accuracy and efficiency, when the WPT is used as a substitute for the DFT in the MFCC method. The novelty of our approach lies in its concentration on the wavelet and the decomposition level as the parameters influencing the performance. We compare the performance of the DFT with the WPT, as well as with our previous work using the DWT. It is shown that the WPT results in significantly lower order for the Gaussian Mixture Model (GMM) used to model speech, and marginal improvement in accuracy with respect to the DFT. WPT mirrors DWT in terms of the order of GMM and can perform as well as the DWT under certain conditions.

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology

Keywords: Cepstral Coefficients, Speaker Recognition, Wavelet Packets ;

* Claude Turner. Tel.: +1-757-823-8311; fax: +1-757-823-9229.
E-mail address: cturner@nsu.edu

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology
doi:10.1016/j.procs.2015.09.177

1. Introduction

Automatic speaker recognition is the identification of a person from his/her voice (Furui, 1997; Campbell,1997; Bimbot et al., 2004). A typical speaker recognition system consists of two phases: an enrollment (or training) phase, and an authentication (or testing) phase. In the enrollment phase, the user speaks an appropriate phrase into a microphone or similar device attached to the system. The system then extracts speaker-specific information from the speech signal in a process called feature extraction. These features are used to build a model for the speaker during the training process. There are many types of models that could be used in speaker recognition, including Gaussian Mixture Models (GMMs) (Reynolds, 1995), Hidden Markov Models, and vector quantization (VQ). However, GMM has been one of the most popular methods for the modeling process. The purpose of the testing phase is to determine whether the speech samples belong to one of the registered speakers. As in the training phase, speech features are extracted from the speech signal presented. The speaker is then determined by finding the speaker model which yields the maximum posterior probability for the input feature vector sequence (Reynolds, 1995).

Feature extraction is the conversion of raw speech signal to acoustic vectors that characterize speaker-specific information. Feature extraction estimates a set of features from the speech signal that represent some speaker-specific information. The speaker-specific information results from complex transformations occurring at multiple levels of the speech production process: semantic, phonologic, phonetic, and acoustic (Atal, 1976;Campbell, 1997). Despite the variation among the categories of speaker-specific information, there are only a small set of criteria that they must satisfy. These are discussed by Nolan and Wolf (Nolan, 2009;Wolf, 1972). There are a variety of filter bank-based feature extraction methods for feature extraction. However, Mel Frequency Cepstral Coefficients (MFCC) (Davis & Mermelstein, 1980) has been the most widely employed approach (Ganchev et al., 2005). In recent years, numerous variations and improvements of the original MFCC idea have been proposed (Ganchev et al., 2005; Sigurdsson et al., 2005). This is mainly attributable to researchers' efforts to exploit progress made in the area of psychoacoustics (Ganchev et al., 2005).

The testing phase of speaker recognition may be cast as a pattern recognition problem. As such, it can be partitioned into two modules (Jin, 2007): (a) a feature extraction module, and (b) a classification module. The feature extraction module is the same as in the training phase. The classification module can be further divided into two components: pattern matching and decision. The *pattern matching* component is responsible for comparing the estimated features to the speaker models. The decision component analyzes the similarity score(s), which could be either statistical or deterministic, to make a decision. The *decision* process is dependent on the system task. For the closed set identification task, the decision could be to select the identity associated with the model that is most similar to the test sample.

The wavelet packet transform (WPT) (or wavelet packet decomposition) has been employed in speaker recognition applications for over two decades with some success (Almaadeed et al., 2015) (Deshpande & Holambe, 2010) (Hsieh et al., 2003) (Sarikaya et al., (1998). Wavelet packets are an extension of the discrete wavelet transform (DWT). The discrete Fourier transform (DFT) is usually employed to compute the spectrum of the speech waveform during the MFCC feature extraction process. However, over the past few years, the discrete wavelet transform (DWT) has gained remarkable attention, and has been favored over the DFT in a wide variety of applications. The DWT enables the decomposition of a signal at multiple layers of resolution. The wavelet packet transform is an extension of the DWT that adds more flexibility to the decomposition process.

This work is a study of the impact on performance in terms of accuracy and efficiency when the WPT is used as a substitute for the DFT in the MFCC feature extraction process. The novelty of this work stems from its exploration of how the use of different wavelets and different decomposition levels in the WPT influences the performance of the speaker identification process. It is shown that the WPT results in significantly lower order for the GMM used to model speaker features and marginal improvement in accuracy. Specifically, we will compare performance in terms of accuracy and efficiency between the DFT and the WPT for Daubechies's first ten wavelets at six different decomposition levels.

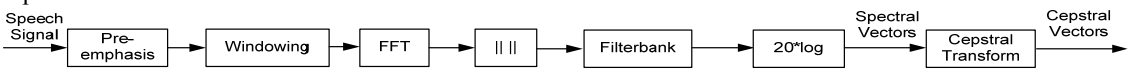

Figure 1 Modular Representation of Feature Extraction

2. Methods

2.1 Mel Frequency Cepstral Coefficients

A modular representation of a filterbank-based feature extraction model that generates the MFCC based feature extraction method is depicted in Fig. 1. The speech signal is first pre-emphasized by applying the filter, $x(t) = y(t) - a \cdot y(t-1)$ where $a \in [0.95, 0.98]$. The goal of the filter is to enhance the high frequencies of the spectrum, which is diminished during the speech production process. Following the pre-emphasis stage is a windowing process, where a window whose size in duration is much smaller than the whole speech signal, is applied starting at the beginning of the signal, and then shifted to the right and applied, successively, until the end of the signal is reached. Two quantities must be set: the width of the window and the shift between consecutive windows. For the width of the window, two values are often used: 20ms and 30ms. These values correspond to the average duration necessary for the stationary assumption to hold. In the case of the delay, a value is chosen so that there is some overlap between consecutive windows. Ten milliseconds is often used.

Once the width of the window and the shift between consecutive windows are found, the type of window can then be chosen. The Hamming and Hanning windows are most often used in speaker recognition. Next, for each of the windowed signals emerging from the windowing process, an N-point DFT is computed. Typically, N is chosen as a power of 2 and is classically 512 points, which is greater than the number of points in the window. Next, the modulus of the DFT for each of the spectral vectors is obtained, and from this the corresponding power spectrum for each is taken over 512 points. Since the signal is real valued, the spectrum is symmetric, thus only the first half plus one sample is kept--257 points. The spectrum consists of much fluctuation. However, in this context, such details are not of interest. It is only the envelope of the spectrum that is of interest. Smoothing removes some of these details. To realize the smoothing and to get the envelope of the spectrum, the spectrum is multiplied by a filterbank. The filterbank is defined by the shape of the filters and by their frequency localization—left frequency, central frequency, and right frequency (Ganchev et al., 2005). Filters may be of a triangular or other shape, and can be differently located on the frequency scale. The Bark/Mel scale is sometimes used for frequency localization of the filters. It is an auditory scale that is similar to the frequency scale of the human ear. A commonly implemented equation for localization of the central frequencies, which is the one used in the experiments of this study, is given by: $\tilde{f} = 1127 \cdot \ln(1 + f_{lin} / 700)$.

The original filterbank of Davis and Mermelstein (Davis & Mermelstein, 1980), FB-20, is the one used here. It proceeds as follows: Given the N-point DFT of the discrete input signal, x, $\hat{x}_k = \sum_{n=0}^{N-1} x_n \exp(-j2\pi nk / N), k \in \{0,1,\cdots,N-1\}$, a filter bank with M equal height triangular filters is constructed. Each of the M equal height filters is defined by:

$$\hat{h}_i(k) = \begin{cases} 0, & k < f_{b_{i-1}} \\ \dfrac{k - f_{b_{i-1}}}{f_{b_i} - f_{b_{i-1}}}, & f_{b_{i-1}} \le k < f_{b_i} \\ \dfrac{f_{b_{i+1}} - k}{f_{b_{i+1}} - f_{b_i}}, & f_{b_i} \le k < f_{b_{i+1}} \\ 0, & k \ge f_{b_{i+1}} \end{cases} \quad (1)$$

$i \in \{1,2,\cdots,M\}$, where i is the filter index, f_{b_i} is the boundary point for the filter, $\hat{h}_i$, and $k \in \{1,2,\cdots,N\}$ corresponds to the *k-th* coefficient of the N-point Discrete Fourier Transform (DFT). Each boundary point, f_{b_i}, depends on the sampling frequency F_s, and the number of points, N, in the DFT, and is given by: $f_{b_i} = N / F_s \cdot \tilde{f}^{-1}(\tilde{f}(f_{low}) + i \cdot (\tilde{f}(f_{high}) - \tilde{f}(f_{low})) / (M+1).$ The values f_{low} and f_{high} are, respectively, the low and high boundary frequencies for the entire filterbank; M, is the number of filters; and $\tilde{f}^{-1}$ is the inverse of (1) given by: $\tilde{f}^{-1} = f_{lin} = 700 \cdot \exp(\tilde{f}_{mel} / 1127 - 1)$, where $f_{low}, f_{high},$ and f_{lin} are in Hertz (Hz), while $\tilde{f}$ is in mels. The filter bank of Davis and Mermelstein is comprised of 20 equal height filters, which cover the frequency range [0,4600] Hz. The center frequencies for the first ten filters are linearly spaced between 100 Hz and 1000 Hz, and the next ten have center frequencies logarithmically spaced between 1000 Hz and 4000 Hz. The next step computes the logarithm of the windowed signal followed by the discrete cosine transform. The process may be summarized compactly as follows:

$$c_t = \sum_{k=1}^{M} X_k \cos\left(t\left(k-1/2\right)\frac{\pi}{M}\right),\ t \in \{1,2,\cdots,T\}, \tag{2}$$

where T is the number of cepstral coefficients computed; usually $T < M$. X_k is referred to as the log energy output of the i-th filter and is given by:

$$X_k = \log_{10}\left(\sum_{k=0}^{N-1} |\hat{x}_i|\hat{h}_k(i)\right). \tag{3}$$

Table 1. Identification Error (out of 6) DWT vs. WPT

Level	Discrete Wavelet Transform										Wavelet Packet Transform									
	db1	db2	db3	db4	db5	db6	db7	db8	db9	db10	db1	db2	db3	db4	db5	db6	db7	db8	db9	db10
1	2	2	2	2	2	2	3	3	3	3	2	2	2	2	2	2	3	3	3	3
2	0	1	1	0	1	0	1	1	1	0	1	1	1	0	1	0	1	1	1	0
3	5	3	4	3	2	3	3	3	3	3	3	3	4	3	2	3	3	3	3	3
4	1	1	2	1	2	2	0	2	4	4	3	1	2	1	2	2	0	2	4	4
5	2	2	3	5	3	4	4	4	5	3	1	2	3	5	3	4	4	4	5	3
6	1	5	3	2	4	2	5	4	3	4	1	5	3	2	4	2	5	4	3	4

2.2 Wavelet Packets

We explain the WPT through the following 3-level WPT decomposition example of the signal, x, of length n. The symbols * and ↓ denote the convolution and downsampling operations, respectively. Sets for the approximation and detail coefficients at a typical level, j, are given by $\{a_x(j,1,\cdot), a_x(j,2),\ldots,a(j,2^{j-1},\cdot)\}$ and $\{d_x(j,1,\cdot), a_x(j,2,\cdot),\ldots,a_x(j,2^{j-1},\cdot)\}$ respectively, where $a_x(j,i,\cdot) \subset \{a_x(j,i,0), a_x(j,i,1),\cdots,a_x(j,i,n_j-1)\}$; $d_x(j,i,\cdot)$ is similarly defined and $a_x(0,\cdot) \subset x$. The number coefficients for a typical subsequence $a_x(j,i,\cdot)$ (or $d_x(j,i,\cdot)$) is denoted n_j. To present a mathematical representation for the WPT, the following sets are introduced: $J = \{1,2,\ldots,L\}, O_j = \{1,3,\ldots,2^j-1\}$, $E_j = \{2,4,\ldots,2^j\}$, $I_j = O_j \cup E_j$ and $K_j = \{0,1,\ldots,n_j-1\}$. Given low- and high pass quadrature mirror filters, h and g, respectively, scaling and wavelet functions, φ and ψ, respectively, and finite signal, x, a mathematical representation for the WPT may be written in the following way:

$$(\forall j \in J)(a_x(j,i,\cdot))(k) = ((a_x(j-1,i-1,\cdot) * h) \downarrow 2)(k) = \sum_{m=0}^{n_{j-1}-1} h_{2k-m} a_x(j-1,i-1,m) \tag{4}$$

$$(\forall j \in J)(d_x(j,i,\cdot))(k) = ((a_x(j-1,i-1,\cdot) * g) \downarrow 2)(k) = \sum_{m=0}^{n_{j-1}-1} g_{2k-m} a_x(j-1,i-1,m) \tag{5}$$

$$(\forall j \in J)(a_x(j,s,\cdot))(k) = ((d_x(j-1,s-1,\cdot) * h) \downarrow 2)(k) = \sum_{m=0}^{n_{j-1}-1} h_{2k-m} d_x(j-1,s-1,m) \tag{6}$$

$$(\forall j \in J)(d_x(j,s,\cdot))(k) = ((d_x(j-1,s-1,\cdot) * g) \downarrow 2)(k) = \sum_{m=0}^{n_{j-1}-1} g_{2k-m} d_x(j-1,s-1,m) \tag{7}$$

$i \in O_j, s \in E_j$. Equations 4 and 5 represent $a_x(j,i,\cdot)$ and $d_x(j,i,\cdot)$ for odd values of i. They are obtained through the approximation coefficients of the previous level, $a_x(j-1,i-1,\cdot)$. Equations 6 and 7 represent $a_x(j,i,\cdot)$ and $d_x(j,i,\cdot)$ for even values of i. They are obtained through the detail coefficients of the previous level, $d_x(j-1,i-1,\cdot)$. We replace the DFT in Fig. 2 by the WPT, and take as its output the following detail signal obtained from the detail coefficients, $d_x(j,i,k)$:

$$D_j(t) = \sum_{i \in I_j} \sum_{k=0}^{n_j-1} d_x(j,i,k)\, \psi_{j,k}(t), \tag{8}$$

To obtain feature extraction filterbank coefficients using the WPT, we substitute the right-side of Eq. 10 for $\hat{x}_k$ in Eq. 5, to get: $z_{kj} = \log_{10}(\sum_{t=0}^{n-1} |D_j(t)|\hat{h}_k(t))$, for some $j \in \{1,2,\cdots,L\}$. The coefficients that results when the WPT is substituted for the DFT in Eq. 3 is then obtained by substituting z_k for X_k in Eq. 2, to get:

$$w_t = \sum_{k=1}^{M} z_{kj} \cos\left(t\left(k-1/2\right)\frac{\pi}{M}\right),\ t \in \{1,2,\cdots,T\} \tag{9}$$

for some $j \in J$.

3. Experimental Setup and Results

We used six Region 1 speakers from the TIMIT database—three males and three females—and the following single utterance from each: “She had your dark suit in greasy wash water all year.” Each speaker has a copy of this

Table 2. Order of GMM: DFT vs. WPT

	Speakers					
	FECD0	FJSP0	FKFB0	MKLS0	MPGH0	MPGR0
DFT	20	20	18	20	20	16
WPT	5	5	5	5	4	5

utterance stored in a file name sa1.wav. The following six speakers were used—three males and three females—from the TIMIT database: FECD0, FJSP0, FKFB0, MKLS0, MPGH0, and MPGR0. The first letter of the speaker designation tells us the gender. The next three letters following it are the first, middle and last initial of the speaker's name. The last position makes it possible to distinguish multiple speakers with the same gender and initials--a zero indicates the first such speaker, a 1 for the second, etc.

In the training phase of our experiment, Eq. 9 was implemented for each of the speaker signals, and a GMM was used to model the features obtained. During the testing phase, for each speaker, Eq. 9 was again used to extract the features, and then a maximum likelihood function was used to determine the model that best matched the input speech. The process was repeated for the first ten Daubechies's wavelets, $\{\text{dbl}, \text{db2}, ..., \text{db10}\}$, and for six decomposition levels of the WPT. These results were compared with the FFT approach given by Eq. 4 and results we obtained when the DWT is applied in a similar manner (Turner et al., 2011). The value of *a* used for the pre-emphasis was $a = 0.95$. The window size, and overlap used in the windowing module was, 320 and 160 samples, respectively. The filterbank was the original filterbank design of Davis and Mermelstein (Davis & Mermelstein, 1980), with 20 filters, $M = 20$, as discussed in Section 2.1. The order of the GMM (number of multivariate Gaussian distributions used) was optimized using the Akaike Information Criteria (Akaike, 1974).

The results for the WPT are provided in Table 1 for the six decomposition level under the column labelled "Wavelet Packet Transform." The results obtained for the DWT in previous study are also provided to the left under the column labelled "Discrete Wavelet Transform." Each value in the table is the total number of speakers that were mis-identified for a particular wavelet at a particular level. To elaborate, a test speaker from the set of speakers previous given is compared against the stored GMM models of all the speakers in this set to determine the closest match. If the speaker is correctly identified, then there is no error contribution to the total. However, if the speaker is mis-identified, the total is increased by 1. This process is repeated for each speaker in the set for each wavelet and for each of the six level. The Table 1 results show that Levels 2 and 6 tied as the best performers for the WPT, with and average performance of 1.4, and db7 on Level 2, and db2 at Level 6 identifying each speaker without error. Level 2 was also observed as the best performer in our previous work with the DWT (Table I). However, we find that the DWT yield superior results in terms of average performance, 0.4, and number of wavelets with zero error. (Wavelets with zero errors at Level for the DWT were db1, db4, db6, and db10, as can be seen in Table I.) The best performance for both the WPT and DWT provide improvement over the DFT, which mis-identified one out of the six speakers (error=1/6).

Table 2 compares GMM order for the WPT versus the DFT. We find that the GMM for the WPT were similar to those obtained for the DWT in (Turner et. al, 2011). The row marked "DFT" gives the number of models used in the training phase for each speaker. There are four speakers, FECD0, FJSP0, MKLS0 and MPGH0, that have order 20. The other two speakers FKFB0 and MPGR0 have order 18 and 16, respectively. In the case of WPT, the results for the GMM order given on the row labeled "WPT," show that the DFT require an order that is three to five times that of the WPT. A smaller optimal order is preferred because it leads to a GMM that is less computationally intensive to generate and use. Therefore, the WPT approach seems to provide marginal improvement over DFT in terms of its accuracy for speaker identification with the MFCC, but provides significant improvement in terms of the optimal order required to generate the GMM.

4. Conclusion

This work compared the performance of the DFT with the WPT in the computation of the MFCC for feature extraction in speaker recognition, when the wavelet and decomposition are used as the parameters. It showed that the speech features derived through the WPT resulted in a more efficient representation, in terms of order, for the GMM that is used in the statistical modeling of features. These results mirrored results we previously obtained for the DWT. It also showed marginal improvement in accuracy of the WPT over the DFT. However, the WPT results on accuracy did not show as much consistency in performance on its best level of performance as the DWT in terms

of average performance and the number of wavelets that yield zero error. It was also shown that the GMM order required when the DFT is used in the MFCC feature extraction process was approximately three to five times that required for the WPT. Finally, in terms of accuracy, we find that the WPT outperforms the DFT in terms of accuracy on db7 at Level 2, and db2 at Level 6.

References

1. Al-Ani, MS, Mohammed, T.S. and Aljebory, K.M., Speaker Identification: A Hybrid Approach Using Neural Networks and Wavelet Transform. *Journal of Computer Science*, 2007, *3*(5), p. 304-309.
2. Almaadeed, N., Aggoun, A., Amira, A., (2015) Speaker identification using multimodal neural networks and wavelet analysis, IET Biom., 2015, 4(1), p. 18–28
3. Akaike, H., A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974, *19*(6), p. 716-723.
4. Atal, B. S. Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, 1976, *64*(4), 460-475.
5. Bimbot, F., Bonastre, J.F., Fredouille, C, Gravieer, G., Magrin-Chagnolleau, I., Meignier, S., et al., A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Applied Signal Processing*, *2004,* 4, 430-451. Hindawi Publishing Corporation.
6. Campbell, J.P., Speaker recognition: a tutorial. *Proceedings of the IEEE*, 1997, *85*(9), 1437-1462.
7. Coifman, R. R. and Wickerhauser, M. V., "Best-adapted wavelet packet bases," Fluid Dynamics Research, 1992, 10, p. 229-250, Accessed at: http://www.math.wustl.edu/~victor/papers/ebafbbs.pdf
8. Davis, S., & Mermelstein, P., Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1980, *28*(4), 357-366.
9. Deshpande, M.S. & Holambe, R.S., "Speaker Identification Using Admissible Wavelet Packet Based Decomposition," World Academy of Science, Engineering and Technology, 2010, 37
10. Furui, S., *Speaker Recognition in Survey of the State of the Art in Human Language Technology*. Cambridge University Press, 1997
11. Ganchev, T., Fakotakis, N., & Kokkinakis, G., Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task. *In Proceedings of the 10th International Conference on Speech and Computer (SPECOM 2005)*, Patras, Greece, 2005, p. 191-194
12. Hsieh, C.-T., Lai, E. and Wang Y.-C., "Robust Speaker Identification System Based on Wavelet Transform and Gaussian Mixture Model," Journal of Information Science and Engineering, 2003, 19, p. 267-282
13. Jin, Q., Robust Speaker Recognition, Carnegie Mellon. 2007, Retrieved from http://www.lti.cs.cmu.edu/Research/Thesis/QinJin.pdf.
14. Nolan, F., *The Phonetic Bases of Speaker Recognition (Cambridge Studies in Speech Science and Communication)*, Cambridge University Press, 2009, p. 5-25
15. Phan, F., Micheli-Tzanakou, E. and Sideman, S., Speaker identification using neural networks and wavelets. *IEEE Engineering in Medicine and Biology Magazine*, 2000, *19*(1), 92-101.
16. Reynolds, D.A., & Rose, R. C., Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 1995, *3*(1), 72-83.
17. Sarikaya, R., Pellom, B.L., Hansen, J.H.L., Wavelet Packet Transform Features with Application to Speaker Identification, Proceedings of IEEE Nordic Signal Processing Symp., Visgo, 1998, p. 81-84
18. Sigurdsson, S., Peterson, K. B., & Lehn-Schiøler, T., Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3 Encoded Music. *Proc. Int. Conf. Music Inf. Retrieval,* 2005, p. 286–289.
19. Siafarikas, M., Ganchev, T., Fakotakis, N., "Wavelet Packet Bases for Speaker Recognition," Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence, Patras, Greece, Oct 29-31, 2007, p. 514-517
20. Turner, C., Joseph, A., Aksu, M. and Langdon, H. (2011) "The Wavelet and Fourier Transform in Feature Extraction for Text-Dependent, Filterbank-Based Speaker Recognition," Procedia Computer Science, 2011, 6, p. 124–129
21. Wolf, J. J., Efficient Acoustic Parameters for Speaker Recognition. *Journal of the American Statistical Association*, 1972, *51*, 2044-2056.