FAST ADAPTIVE COMPONENT WEIGHTED CEPSTRUM POLE FILTERING FOR SPEAKER IDENTIFICATION

Arthur L. Swanson[1], *Ravi P. Ramachandran*[2] *and Steven H. Chin*[3]

1. L-3 Communications, arthur.l.swanson@l-3com.com
2. Rowan University, ravi@rowan.edu
3. Rowan University, chin@rowan.edu

ABSTRACT

Mismatched training and testing conditions for speaker identification exist when speech is subjected to a different channel for the two cases. This results in diminished speaker identification performance. Finding features that show little variability to the filtering effect of different channels will make speaker identification systems more robust thereby achieving a better performance. It has been shown that subtracting the mean of the pole filtered linear predictive (LP) cepstrum from the actual LP cepstrum results in a robust feature. This feature is known as the pole filtered mean removed LP cepstrum. Another robust feature is the adaptive component weighted (ACW) cepstrum particularly with mean removal. In this paper, we combine the ACW cepstrum with the pole filtering concept to configure a more robust new feature, namely, the pole filtered mean removed ACW cepstrum. This new method is fast and shows a higher performance then the pole filtered mean removed LP cepstrum and the mean removed ACW cepstrum. Experimental results are given for the TIMIT database involving a variety of mismatched conditions.

1. INTRODUCTION

Speaker recognition refers to the concept of recognizing a speaker by his/her voice or speech samples [1][2][3]. Some of the important applications of speaker recognition include customer verification for bank transactions, access to bank accounts through telephones, control on the use of credit cards, and for security purposes in the army, navy and airforce. The two tasks within speaker recognition are *speaker identification* and *speaker verification*. Speaker identification (ID) deals with a situation where the person has to be identified as being one among a set of persons by using his/her voice samples. The objective of speaker verification is to verify the claimed identity of that speaker based on the voice samples of that speaker alone. A claimant speaker is either accepted or rejected by the system.

The speaker ID problem may further be subdivided into *closed set* and *open set*. The closed set speaker ID problem refers to a case where the speaker is known *a priori* to belong to a set of M speakers. In the open set case, the speaker may be out of the set and hence, a "none of the above" category is necessary. Another distinguishing aspect of speaker recognition systems is that they can either be text-independent or text-dependent depending on the application. In the text-independent case, there is no restriction on the sentence or phrase to be spoken, whereas in the text-dependent case, the input sentence or phrase is fixed for each speaker. The focus of this paper is on text-independent, closed set speaker identification.

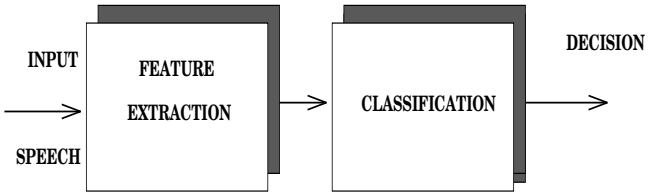

Figure 1: A general diagram of a recognition system

Speaker recognition consists of two stages, namely, *Feature Extraction* and *Classification* as shown in Fig. 1. Feature extraction is associated with obtaining the characteristic patterns of the signal that are representative of the speaker in question. The parameters or features used in speaker recognition are a transformation of the speech signal into a compact acoustic representation that contains information useful for the identification of the speaker. This is often done using short-time linear predictive (LP) [4] analysis which leads to an all-pole LP vocal tract model. The LP coefficients are converted to the LP cepstrum [4] which in turn, is the feature vector. The classifier uses the features to render a decision as to the speaker identity or verifies the claimed identity of the speaker.

The recognition task is highly successful if the environmental conditions for training and testing are the same (known as matched conditions). Studies have shown that recognition performance degrades when the training and testing conditions are not the same (known as mismatched conditions) [5][6][7]. This occurs when the speaker is trained on one type of telephone (handset, cordless or speakerphone) and during the testing phase, a different type of telephone is used. In this particular case, channel mismatch is encountered and this contributes to the degradation in the performance. Channels have a filtering effect on the speech and alter the overall spectral envelope of the speech signal. Assuming that the speech and channel spectra are well approximated by the all-pole LP model, it is observed that a channel influence on the speech leads to an additive component on the LP cepstrum. Estimating and removing this additive channel component will mitigate the channel effect and make speaker recognition systems more robust. One method of estimating the additive channel component is to take the mean of the LP cepstrum vectors over an utterance [8]. It has been shown that a better estimate is obtained by taking the mean of the pole filtered LP cepstrum [9][10][11]. Removal of the mean of the pole filtered LP cepstrum from the LP cepstrum vectors results in a more

robust feature and is known as the pole filtered mean removed cepstrum (PFMRC). Another channel estimate is based on the mean of the adaptive component weighted (ACW) cepstrum [12][13][14]. Removal of the mean of the ACW cepstrum from the ACW cepstrum vectors results in another robust feature and is known as the mean removed ACW cepstrum (MRACW).

In this paper, we combine the concept of pole filtering to the ACW cepstrum to configure a new channel estimate and a new feature. The channel estimate is the pole filtered mean of the ACW cepstrum. Removal of the pole filtered mean of the ACW cepstrum from the ACW cepstrum vectors results in the new robust feature and is known as the pole filtered mean removed ACW cepstrum (PFMRACW). This method is computationally efficient just like its PFMRC and MRACW counterparts. It also gives a better performance than the PFMRC and MRACW approaches.

2. LINEAR PREDICTIVE CEPSTRUM

The autoregressive LP model for speech is given by the difference equation [4]

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + e(n) \tag{1}$$

where $s(n)$ is the speech signal, $e(n)$ is the prediction error and a_i are the predictor coefficients. It can be noted that $s(n)$ is predicted as a linear combination of the previous p samples. The all-pole LP transfer function is given by

$$H(z) = \frac{S(z)}{E(z)} = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}}. \tag{2}$$

where $S(z)$ and $E(z)$ are the z-transforms of $s(n)$ and $e(n)$ respectively. In practice, the predictor coefficients a_i are computed over short intervals (typically 10 ms to 30 ms) called frames during which the vocal tract configuration is assumed to be stationary. This is done using the autocorrelation method [4][15] which guarantees that $H(z)$ is a stable function.

The predictor coefficients a_i are converted to the LP cepstrum $clp(n)$ $(n \geq 1)$ by an efficient recursive relation given as [4][15]

$$clp(n) = a_n + \sum_{i=1}^{n-1} (\frac{i}{n}) clp(i) a_{n-i} \tag{3}$$

Since $clp(n)$ is of infinite duration, the feature vector of dimension p consists of the components $clp(1)$ to $clp(p)$ which are the most significant due to the decay of the sequence with increasing n.

2.1. Pole Filtered Mean Removed Cepstrum (PFMRC)

As mentioned earlier, when speech is subjected to channel interference, an additive component due to the channel manifests itself on the LP cepstrum. To compensate for the channel effect, this component is estimated as the mean or the pole filtered mean of the LP cepstrum and removed by subtraction. For simple mean subtraction, the feature vector is

$$cmrc(n) = clp(n) - E[clp(n)] \tag{4}$$

where the expectation is taken over an utterance consisting of a number of frames.

The LP poles with narrow bandwidths that lie close to the unit circle usually represent the formants and are less sensitive to channel and noise effects. Hence, these poles do not contribute to the channel estimate as they contain much speech information. In contrast, the broad bandwidth poles model the spectral tilt, sub-glottal variation and the channel effects. These poles offer a better estimate of the channel. Pole filtering modifies the LP poles so as to broaden the bandwidth of the formant poles [9][10][11]. Broadening the bandwidth of the formant poles is performed by transforming the LP polynomial so as to weight the predictor coefficients as given by

$$H(z/\gamma) = \frac{1}{A(z/\gamma)} = \frac{1}{1 - \sum_{i=1}^{p} a_i \gamma^i z^{-i}}. \tag{5}$$

where $0 < \gamma \leq 1$. Given the original LP poles p_i, the new set of poles are γp_i. The cepstrum formed from these modified poles (denoted as $cpflp(n)$) is related to the LP cepstrum as [9][10][11].

$$cpflp(n) = \gamma^n clp(n) \tag{6}$$

The feature vector $cpfmrc(n)$ is known as the pole filtered mean removed cepstrum (PFMRC) and is computed as given below.

- Select the parameter γ.
- For each frame of speech, calculate $clp(n)$ and $cpflp(n)$.
- Find the channel estimate $E[cpflp(n)]$ where the expectation is taken over all speech frames in an utterance.
- Find the feature vector $cpfmrc(n) = clp(n) - E[cpflp(n)]$.

3. ADAPTIVE COMPONENT WEIGHTED CEPSTRUM

The first step in developing the ACW cepstrum [12] is to perform a partial fraction expansion of the LP function $H(z) = 1/A(z)$ to get

$$\begin{aligned} \frac{1}{A(z)} &= \sum_{k=1}^{p} \frac{\lim_{z \to p_k} \left[\left(1 - p_k z^{-1}\right) / A(z) \right]}{1 - p_k z^{-1}} \\ &= \sum_{k=1}^{p} \frac{r_k}{1 - p_k z^{-1}} \end{aligned} \tag{7}$$

The experiments in [12] reveal that the residues r_k show considerable variations especially for nonformant poles when the speech is degraded. Therefore, the variations in r_k were removed by forcing r_k to be $constant = 1$ for every k. Hence, the resulting transfer function is a pole-zero type of the form

$$\begin{aligned} H_{acw}(z) &= \frac{N(z)}{A(z)} \\ &= \sum_{k=1}^{p} \frac{1}{1 - p_k z^{-1}} \\ &= \frac{1}{A(z)} \sum_{k=1}^{p} \prod_{i=1 \neq k}^{p} (1 - p_i z^{-1}) \\ &= p \frac{1 - \sum_{k=1}^{p-1} b_k z^{-k}}{1 - \sum_{k=1}^{p} a_k z^{-k}} \end{aligned} \tag{8}$$

It has been shown in [14] that $N(z)$ is the derivative of $A(z)$ with respect to z and hence, the coefficients b_k are easily found from a_k. Applying the recursion in Eq. (3) to b_k and a_k results in two cepstrum sequences $cb(n)$ and $clp(n)$ respectively. The ACW cepstrum is $cacw(n) = clp(n) - cb(n)$. For simple mean subtraction (MRACW method), the feature vector is

$$cmracw(n) = cacw(n) - E[cacw(n)] \tag{9}$$

where the expectation is taken over an utterance consisting of a number of frames.

3.1. Pole Filtered Mean Removed Adaptive Component Weighted Cepstrum (PFMRACW)

The contribution of this paper is to combine the pole filtering concept to the ACW cepstrum to get a better channel estimate and a more robust feature vector. The first step is to choose a value of γ between 0 and 1 and perform a partial fraction expansion of the pole filtered LP function $1/A(z/\gamma)$ to get

$$\frac{1}{A(z/\gamma)} = \sum_{k=1}^{p} \frac{s_k}{1 - q_k z^{-1}} \tag{10}$$

Setting $s_k = 1$ for every k gives a transfer function

$$\begin{aligned} H_{pfacw}(z) &= \frac{M(z)}{A(z/\gamma)} \\ &= \sum_{k=1}^{p} \frac{1}{1 - q_k z^{-1}} \\ &= p \frac{1 - \sum_{k=1}^{p-1} m_k z^{-k}}{1 - \sum_{k=1}^{p} \gamma^k a_k z^{-k}}, \end{aligned} \tag{11}$$

Again, $M(z)$ is the derivative of $A(z/\gamma)$ with respect to z and hence, the coefficients m_k are easily found from $\gamma^k a_k$. Applying the recursion in Eq. (3) to m_k results in the cepstrum sequence $cm(n)$. The cepstrum corresponding to the denominator of $H_{pfacw}(z)$ is $cpflp(n)$ (see Eq. (6). The pole filtered ACW cepstrum is expressed as $cpfacw(n) = cpflp(n) - cm(n)$. The feature vector (denoted as $cpfmracw(n)$) is known as the pole filtered mean removed ACW cepstrum (PFMRACW) and is computed as given below.

- Select the parameter γ.
- For each frame of speech, calculate $cacw(n)$ and $cpfacw(n)$.
- Find the channel estimate $E[cpfacw(n)]$ where the expectation is taken over all speech frames in an utterance.
- Find the feature vector $cpfmracw(n) = cacw(n) - E[cpfacw(n)]$.

4. EXPERIMENTAL RESULTS

Closed set, text-independent speaker identification experiments are carried out using the TIMIT database. Thirty eight speakers from the New England dialect are considered. The speech is downsampled from 16 kHz to 8 kHz. For each speaker, there are 10 sentences. The first five are used for training a vector quantizer (VQ) classifier using the Linde-Buzo-Gray (LBG) method [16] and the squared Euclidean distance as the distortion measure. A VQ codebook is designed for each of the 38 speakers. The training conditions include clean speech and speech subjected to representative bandpass telephone channels [17]: (1) the Continental Mid Voice (CMV) channel, (2) the Continental Poor Voice (CPV) channel, (3) the European Mid Voice (EMV) channel and (4) the European Poor Voice (EPV) channel. The remaining five sentences are individually used for testing thereby giving 190 test utterances.

Training Condition	Testing Condition	PFMRC γ, ISR	MRACW ISR	PFMRACW γ, ISR
Clean	CMV	0.90, 58.4	60.0	0.95, 65.8
Clean	CPV	0.90, 59.5	56.3	0.90, 67.4
CMV	CPV	0.85, 66.3	70.0	0.80, 74.2
CPV	CMV	0.80, 66.8	65.8	0.90, 73.7
Clean	EMV	0.70, 67.9	65.3	0.80, 79.5
Clean	EPV	0.75, 70.5	68.4	0.70, 78.4
EMV	EPV	0.80, 74.2	70.0	0.85, 81.6
EPV	EMV	0.70, 71.6	70.5	0.90, 78.4
CMV	EMV	0.80, 64.2	62.6	0.95, 68.4
CMV	EPV	0.95, 62.6	62.6	0.95, 66.3
CPV	EMV	0.85, 58.9	55.8	0.75, 62.1
CPV	EPV	0.85, 61.6	57.4	0.95, 58.9

Table 1: Speaker identification success rate (ISR) as a percent. The best value of γ is shown. The acronym PFMRC is for pole filtered cepstral mean subtraction. The acronym MRACW is for adaptive component weighted cepstral mean subtraction. The acronym PFMRACW is for pole filtered adaptive component weighted cepstral mean subtraction.

The testing conditions correspond to channel corrupted speech. Consider a particular test feature vector. This is quantized by each of the 38 codebooks. The quantized vector is that which is closest (according to the squared Euclidean distance) to the test feature vector. Hence, 38 different distances are recorded, one for each codebook. This process is repeated for every test feature vector. The distances are accumulated over the entire set of feature vectors. The codebook which renders the smallest accumulated distance identifies the speaker. The identification success rate (ISR) is the number of utterances for which the speaker is identified correctly divided by the total number of utterances tested. The VQ codebook size is 64. For both training and testing, silent or low-energy frames are discarded by energy thresholding. Also, a 12th order LP analysis is used with 30 ms frames having an overlap of 20 ms. All the feature vectors have dimension 12.

Table 1 shows the results for speaker identification. Different values of γ were tried. The best values of γ are from 0.70 to 0.95. In Table 1, the result for the best value of γ is given for each training/testing combination. The performance of our new PFMRACW method is better than the both the PFMRC and MRACW approaches (except for only one case when the CPV channel is used for training and the EPV channel is used for testing).

For both the PFMRC and PFMRACW methods, the best value of γ depends on the training and testing conditions. However, results show that the ISR varies very little for values of γ between 0.70 and 0.95. Decreasing γ below 0.70 does result in significant performance loss and hence, these values should not be used. Figure 2 shows the ISR versus γ for the PFMRC and PFMRACW methods for the case when training is done on the CMV channel and testing is done on the CPV channel. The question of what γ to use can more easily be answered since the variation in the ISR is relatively low for values of $0.70 \leq \gamma \leq 0.95$. By examining the results for

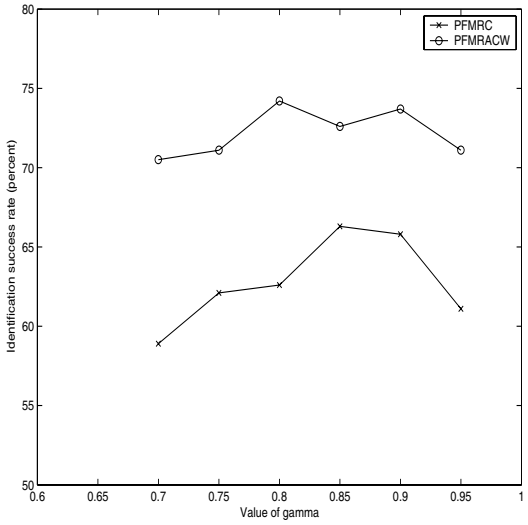

Figure 2: The ISR versus γ for the PFMRC and PFMRACW methods for the case when training is done on the CMV channel and testing is done on the CPV channel.

Training Condition	Testing Condition	PFMRC ISR	MRACW ISR	PFMRACW ISR
Clean	CMV	56.3	60.0	56.3
Clean	CPV	57.4	56.3	60.5
CMV	CPV	66.3	70.0	72.6
CPV	CMV	63.2	65.8	73.7
Clean	EMV	62.6	65.3	77.9
Clean	EPV	68.9	68.4	77.9
EMV	EPV	71.6	70.0	81.6
EPV	EMV	70.5	70.5	76.8
CMV	EMV	56.8	62.6	62.6
CMV	EPV	61.1	62.6	51.1
CPV	EMV	58.9	55.8	59.5
CPV	EPV	61.6	57.4	57.4

Table 2: Speaker identification success rate (ISR) as a percent for $\gamma = 0.85$ is shown. The acronym PFMRC is for pole filtered cepstral mean subtraction. The acronym MRACW is for adaptive component weighted cepstral mean subtraction. The acronym PFMRACW is for pole filtered adaptive component weighted cepstral mean subtraction.

all the training and testing conditions attempted, it is beneficial to fix γ at 0.85. Table 2 shows the results for $\gamma = 0.85$. It is clear that the ISR for both the PFMRC and PFMRACW are generally slightly below the best possible and that the PFMRACW method is almost always the best (the exception being two cases). Table 2 also shows the results for the MRACW method for the sake of completeness.

5. REFERENCES

1. G. R. Doddington, "Speaker recognition - identifying people by their voices" *Proc. IEEE*, vol. 73, pp. 1651–1664, November 1985.
2. A. E. Rosenberg, "Automatic speaker verification: A review", *Proc. IEEE*, vol. 64, pp. 475–487, April 1976.
3. J. P. Campbell, "Speaker recognition: A tutorial", *Proc. IEEE*, vol. 85, pp. 1437–1462, September 1997.
4. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.
5. A. E. Rosenberg and F. K. Soong, "Recent research in automatic speaker recognition", in *Advances in Speech Signal Processing*, edited by S. Furui and M. M. Sondhi, Marcel Dekker, pp. 701–738, 1991.
6. R. J. Mammone, X. Zhang and R. P. Ramachandran, "Robust speaker recognition - A feature based approach", *IEEE Signal Proc. Mag.*, vol. 13, pp. 58–71, September 1996.
7. R. P. Ramachandran, K. R. Farrell, Roopashri Ramachandran and R. J. Mammone, "Robust Speaker Recognition - General Classifier Approaches and Data Fusion Methods", *Pattern Recognition*, Vol. 35, No. 12, pp. 2801–2821, December 2002.
8. B. S. Atal, ""Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *Jour. of the Acoust. Soc. of Amer.*, vol. 55, pp. 1304–1312, June 1974.
9. D. Naik and R. J. Mammone, "Channel normalization using pole filtered cepstral mean subtraction", *SPIE Int. Symp. on Optics, Imaging and Instrumentation*, vol. 2277, San Diego, California, pp. 99–110, July 1994.
10. D. Naik and R. J. Mammone, "Pole filtered cepstral mean subtraction", *IEEE Int. Conf. on Acoust., Speech and Sig. Proc.*, Detroit, Michigan, pp. 157–160, May 1995.
11. R. P. Ramachandran and K. R. Farrell, "Fast Pole Filtering for Speaker Recognition", *IEEE Int. Symp. on Circuits and Systems*, Geneva, Switzerland, pp. V-49–V-52, May 28–31, 2000.
12. K. T. Assaleh and R. J. Mammone, "New LP-derived features for speaker identification", *IEEE Trans. on Speech and Audio Proc.*, vol. 2, pp. 630–638, October 1994.
13. M. S. Zilovic, R. P. Ramachandran and R. J. Mammone, "Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions", *IEEE Trans. on Speech and Audio Proc.*. vol. 6, pp. 260–267, May 1998.
14. M. S. Zilovic, R. P. Ramachandran and R. J. Mammone, "A Fast Algorithm for Finding the Adaptive Component Weighted Cepstrum for Speaker Recognition", *IEEE Trans. on Speech and Audio Processing*, Vol. 5, No. 1, pp. 84–86, January 1997.
15. L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.
16. Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design", *IEEE Trans. on Comm.*, vol. COM-28, pp. 84–95, Jan. 1980.
17. J. Kupin, "A wireless simulator (software)," CCR-P, April 1993.