Highly Noise Robust Text-dependent Speaker Recognition based on Hypothesized Wiener Filtering

V. Ramasubramanian, Deepak Vijaywargiay*, V. Praveen Kumar

Siemens Corporate Technology - India
Siemens Information Systems Ltd., Bangalore - 560100, India
{V.Ramasubramanian, V.Praveenkumar}@siemens.com, Deepak.Vijaywargiay@iiitb.ac.in

Abstract

We propose a hypothesized Wiener filtering (HWF) algorithm for noise robust variable-text text-dependent speaker-recognition. The proposed algorithm exploits an important feature of the text - dependent mode of operation of speaker-recognition, namely, the availability of the 'clean reference templates' of the words of the 'password' text which is supposed to be the text of the input noisy speech. The proposed HWF algorithm is set within the one-pass DP framework proposed by us recently for text-dependent speaker-recognition, which enables use of multiple-templates for each word in the password. We evaluate the proposed HWF algorithm for both speaker - identification and speaker - verification using the TIDIGITS database and show that the proposed HWF algorithm has very high recognition accuracies for both additive white-noise conditions and non-stationary color noise conditions (factory, chopper and babble noises), which are also the typical conditions where conventional spectral subtraction techniques perform poorly.
Index Terms: Robust speaker recognition, hypothesized Wiener filtering, text-dependent speaker recognition, one-pass DP algorithm

1. Introduction

Robustness of a speaker-recognition system to additive background noise is an important problem when the system needs to operate in noisy environments. This is an even more challenging task when the system has to perform recognition in a noisy environment different from that of training. This is typically the case for speaker-recognition applications such as access control to buildings, cars, offices etc., or speaker-authentication over telephones / mobiles (prior to secure tele-transactions) where a high degree of background noise in the form of street noise, car noise, other people's speech (babble noise) etc., can be expected.

A conventional approach to dealing with noisy speech in applications such as speech recognition, text-independent speaker-recognition and speech coding is to apply noise-removal techniques such as spectral-subtraction or conventional Wiener filtering methods so as to get an enhanced speech signal prior to feature extraction. While spectral subtraction requires an estimate of the noise power spectral densities, typically from the most recent non-speech region, Wiener filtering methods require estimates of both clean speech power spectrum and the noise power spectrum. There are a wide variety of Wiener filtering techniques depending on how the clean speech power spectrum estimate is obtained for any given frame; these can be broadly categorized as based on spectral-subtraction or, estimates of signal spectrum from previous 'cleaned' frames or, from model-based estimates such as linear-prediction, or using vector quantizer codebooks; these methods are typically employed in an iterative framework [1].

The hypothesized Wiener filtering (HWF) [2] was originally proposed for robust speaker-dependent isolated word recognition in a DTW framework and has subsequently been adapted to HMM frameworks using state-based filtering for noisy speech recognition [3], [4]. However, despite its appealing feature of making use of clean templates or HMMs, it has not been used for speaker-recognition applications so far, possibly due to the larger focus of research on text-independent speaker-recognition, where HWF cannot be applied [5].

In this paper, we propose the use of the hypothesized Wiener filtering (HWF) approach for realizing a noise robust variable-text text-dependent speaker-recognition system. The 'text-dependent' speaker-recognition problem represents an unique and ideal setting for deriving an advantage with the HWF algorithm, wherein 'clean reference templates' of the words of the 'password' text which is supposed to be the text of the input noisy speech are available. The proposed HWF approach exploits this effectively for robust speaker-recognition with high recognition accuracies for both additive white-noise and non-stationary color noise conditions. The proposed HWF algorithm is set within the one-pass dynamic programming (DP) framework proposed by us recently for variable-text text-dependent speaker-recognition [6], [7], which enables use of multiple-templates for each word in the password so as to capture the intra-speaker variabilities adequately.

2. One-pass DP based speaker-recognition

Fig. 1 shows the typical architecture of the variable-text speaker-recognition system based on the one-pass dynamic programming (DP) matching algorithm proposed by us recently [6], [7]. Here, the figure shows the matching for one speaker; each speaker has a set of templates for each word in the vocabulary. Given an input utterance, the feature extraction module converts the utterance into a sequence of feature vectors (such as the mel-frequency-cepstral coefficients (MFCCs)). This feature vector sequence corresponds to the input 'password' text (say, the digit string 915 in the figure).

The one-pass DP algorithm matches the input feature-vector sequence against the word-models of 9 1 and 5, using multiple templates per word and inter-word silence templates. The resulting match score (D_k^*) is the optimal distance between the input utterance and the word-templates of speaker S_k. For closed-set speaker-identification, this score is computed for each speaker and

* On internship from International Institute of Information Technology, Bangalore, India

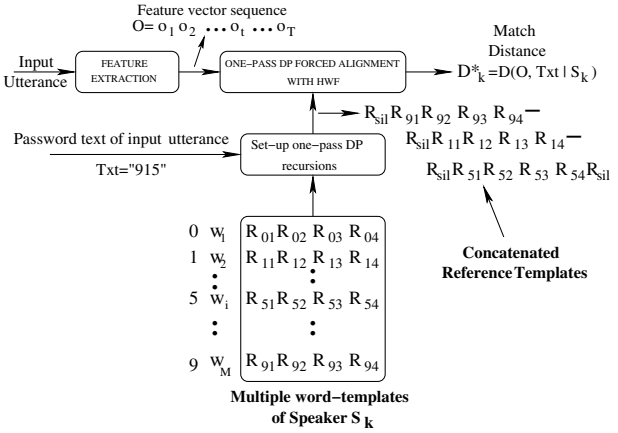

Figure 1: *Text-dependent speaker recognition using Hypothesized Wiener Filtering (HWF) and One-pass DP framework*

the speaker with the lowest score is declared the input speaker. For speaker-verification, this score corresponds to the match between the input utterance and the claimed speaker S_k's models; this score is normalized by the background score, computed between the input utterance and background speaker's word-templates and the normalized score is compared to a threshold; the input speaker claim is accepted if the normalized score is less than the threshold and rejected otherwise.

3. Proposed HWF based speaker-recognition

Fig. 2 shows how the input utterance on the x-axis (corresponding to the password text '915') is matched against the reference templates 'R_9 R_1 R_5' on the y-axis for a speaker. This is a simplified illustration of the one-pass DP matching which in actuality uses multiple templates of each word in the password text so that any of the multiple template of each word is selected for the optimal match and also uses inter-word silence templates so as to allow for inter-word pauses to be present or absent in the input utterance. This algorithm has been described in detail in [6] and [7].

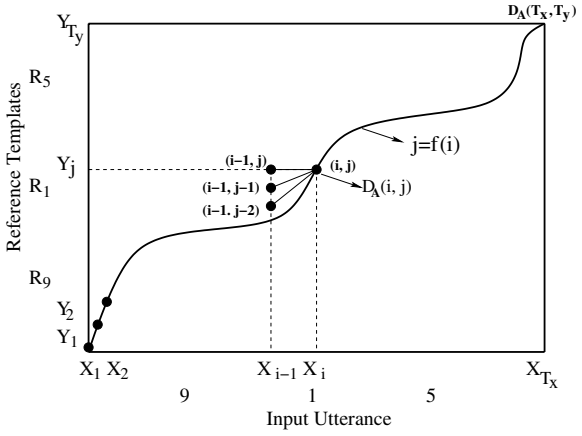

Figure 2: *Dynamic time-warping (DTW) path for HWF*

The HWF based algorithm proposed here is described as follows. The input speech (x-axis) is represented by T_x frames $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_{T_x}$, where $\mathbf{x}_i$ is the sequence of speech samples of the i^{th} frame and the corresponding sequence of MFCC feature vectors is $X_1, X_2, \ldots, X_i, \ldots, X_{T_x}$. The sequence of MFCC feature vectors for the concatenated reference templates on the y-axis is $Y_1, Y_2, \ldots, Y_j, \ldots, Y_{T_y}$.

Let the power spectral density (psd) of the input speech be $P_{\mathbf{x}_1}(w), P_{\mathbf{x}_2}(w), \ldots, P_{\mathbf{x}_i}(w), \ldots, P_{\mathbf{x}_{T_x}}(w)$ and the psd of the concatenated reference templates be $P_{\mathbf{y}_1}(w), P_{\mathbf{y}_2}(w), \ldots, P_{\mathbf{y}_j}(w), \ldots, P_{\mathbf{y}_{T_y}}(w)$. Let $P_{\mathbf{n}}(w)$ be the noise-estimate obtained from the most recent non-speech region of the input noisy speech.

The DTW matching is an optimal time-alignment between the input utterance and the reference templates wherein the warping function $j = f(i)$ relates the i-th frame of the input utterance to the j-th frame of the reference templates such that the accumulated distance between frame-i of input utterance and frame-j of reference templates over the warping path is minimized. The point (i, j) is associated with the minimum accumulated distortion $D_A(i, j)$ which is given by the recursion (as shown in the figure)

$$D_A(i,j) = \min_{k \in \{j, j-1, j-2\}} [D_A(i-1,k) + d_w(i,j)]$$

$$d_w(i,j) = d(\tilde{X}_{ij}, Y_j)$$

where $d(\tilde{X}_{ij}, Y_j)$ is the Euclidean distance between the MFCC vectors $\tilde{X}_{ij}$ and Y_j. While Y_j is the MFCC vector of the j^{th} frame of the concatenated reference template, $\tilde{X}_{ij}$ is the MFCC vector of the speech signal of the i^{th} frame given by $\tilde{\mathbf{x}}_{ij}$ obtained by Wiener filtering the input noisy frame $\mathbf{x}_i$ using the Wiener filter frequency response given by

$$W_j(w) = \frac{P_{\mathbf{y}_j}(w)}{P_{\mathbf{y}_j}(w) + P_{\mathbf{n}}(w)}$$

This is done by computing the psd $P_{\tilde{\mathbf{x}}_{ij}}(w) = P_{\mathbf{x}_i}(w) \cdot W_j(w)$ and obtaining $\tilde{\mathbf{x}}_{ij}$ as a frame of time domain samples corresponding to the psd $P_{\tilde{\mathbf{x}}_{ij}}(w)$. The MFCC vector $\tilde{X}_{ij}$ is then obtained from $\tilde{\mathbf{x}}_{ij}$.

By this, the DTW (or the one-pass DP algorithm [6]) is computed on a grid of local distances $d_w(i,j), i = 1, \ldots, T_x, j = 1, \ldots, T_y$, where each column i contains the local distances between the 'clean' frames $y_j, j = 1, \ldots, T_y$ and the corresponding 'cleaned' frames $\tilde{\mathbf{x}}_{ij}, j = 1, \ldots, T_y$. Clearly, since the reference template is the 'clean' version of the input noisy speech, a column i will have the lowest $d_w(i,j)$ for that j which corresponds to the noisy frame i within a non-linear warping factor. Thus the DTW will now find the optimal warping path $j = f^*(i)$ which minimizes the accumulated distortion $D^* = D_A(T_x, T_y)$ given by

$$D^* = \min_{f(i)} \sum_{i=1}^{T_x} d_w(i, f(i))$$

This D^* corresponds to the HWF score D_k^* for speaker k, i.e., when the y-axis of Fig. 2 uses the speaker k 'clean' templates. Speaker-identification and speaker-verification are done using the score D_k^* as described earlier in Sec. 2 (Fig. 1).

The important point to note here is that D_k^* will be the lowest for the correct speaker, though for each speaker k, the y-axis in Fig. 2 uses the 'same password', (i.e., $R_9 R_1 R_5$ for the example shown) but which are speaker k's word templates. Here, the HWF exploits the fact that the correct speaker's reference templates provide a better 'cleaning' up of the input noisy speech (x-axis) and hence in lower local distances along the optimal path for that speaker than for other speakers.

This is illustrated in Fig. 3, where the matrix of local distances $d_w(i,j), i = 1, \ldots, T_x, j = 1, \ldots, T_y$ is plotted for a single digit (digit 8) test utterance for the three cases: i) No HWF is performed,

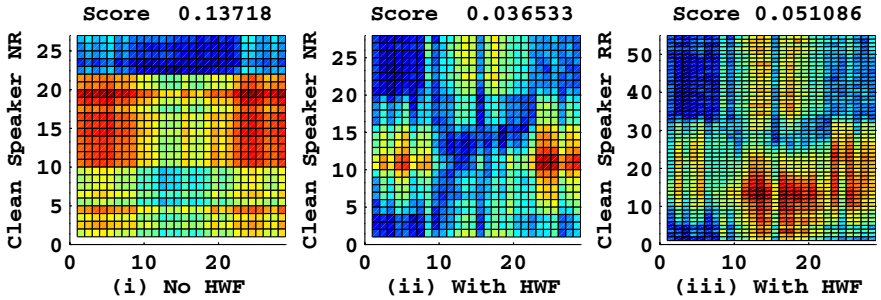

Figure 3: *DTW local distance matrix between noisy input speech (x-axis) and clean templates (y-axis) for i) No HWF with same speakers (NR), ii) with HWF for same speakers (NR), iii) with HWF for different speakers (NR and RR);* **[blue: low values, red: high values]**

but y-axis with clean reference template of the same speaker as the input noisy speech (NR), ii) HWF performed with y-axis having the clean reference template of the same speaker as the input noisy speech (NR), iii) HWF performed with y-axis having the clean reference template of a speaker (RR) who is **not** the speaker of the input noisy speech (NR). Clearly, it can be noted that case-(i) performs poorly though the same speaker template is used; the conventional DTW (without HWF) is unable to find any good optimal path for matching and this results in the loss of recognition accuracy as the input speaker becomes confusable with other speakers. Case-(ii) shows that HWF is able to find a good optimal path after using local distances derived after the HWF operation on the noisy speech using the correct speaker's clean template. Case-(iii) again performs poorly with no good optimal path as there is a mismatch between the speaker of the input speech and the reference templates. HWF actually adds to the discriminability by ensuring that the local distances resulting from case-(iii) are higher than in case-(ii), due to the fact that, in case-(iii), the noisy input frames are filtered by 'clean' spectra of some other speaker (though the textual content is same). The figure also shows the matching score D^* for each of these case, clearly validating the above differences.

It should be noted that this is a far more demanding requirement than the isolated word recognition (IWR) task on which HWF was originally proposed for (and has been used so far) [2], [3], [4]. In the case of using HWF for IWR, the 'correct word' naturally provides a better 'cleaning' up and hence a lower DTW score than an 'incorrect word' whose spectral content obviously does not match the input speech. In contrast, the HWF's task in speaker-recognition is all the more difficult since, for a given speaker, the DTW-HWF algorithm needs to provide a better match only when there is a 'speaker match', despite having the same 'word-content' between the x-axis and y-axis for all speakers.

4. Experiments and Results

We now present results of the HWF algorithm proposed here, used within the one-pass DP framework [6] for text-dependent speaker recognition. We evaluate the HWF algorithm proposed here for both speaker-identification (closed-set) and speaker-verification on 8 speakers in the TIDIGITS database which has a 11 word vocabulary {'oh', 0-9}. The clean templates were extracted from the 7-digits strings and the test data consists of 3, 4, and 5 digit strings with 11 utterances each per speaker. These experiments adequately bring out the basic performance potential of HWF. We also compare it with conventional spectral-subtraction [8], [1] of the input noisy speech before feature extraction in Fig. 1. These algorithms (along with the baseline performance of 'noisy speech' without any noise-removal) are evaluated for clean test data and for additive white-noise of SNRs 0 dB, 5 dB and 10 dB and non-stationary noises (factory, chopper and babble) of 0 dB SNR (from NOISEX92 database).

We performed comparisons with spectral subtraction to bring out an important difference between spectral subtraction and Wiener filtering. Spectral subtraction depends solely on obtaining a noise - estimate (from the most recent non-speech region) and subtracting it from successive speech spectra and then generating the enhanced speech by overlap-add-synthesis method. The performance of spectral subtraction therefore depends only on how well the noise-estimate matches the noise spectra of the noisy speech so that spectral subtraction removes the noise. In fact, we have observed this to work quite well when the noise is stationary colored noise, such as car-noise, which allows the spectral subtraction to effectively subtract out the color noise spectra from the noisy speech regions using the noise-estimate which correctly has the form of a spectral envelope of the stationary color noise [6], [7].

However, when the input noise is white noise (or non-stationary color noise), the spectral subtraction technique fails completely since the noise-estimate obtained from one non-speech region no longer matches the white noise spectra in a subsequent noisy speech region. The white noise estimate exhibits random spectral variations about a flat spectral envelope and therefore does not subtract out a similar flat, but equally random white noise spectra in a noisy speech region. Thus, subtraction actually leaves behind a remnant white noise spectra and at best (when the noise-estimate becomes more and more flat due to longer time-averages in the non-speech region) results in the original noisy speech spectra to have a reduced spectral average, equivalent to an overall attenuation of the noisy speech without any enhancement; the resultant speech therefore provides no improved recognition accuracy.

Spectral subtraction behaves in a similar way for non-stationary color noise also, where the color noise spectra is time varying and the noise-estimate used by the spectral subtraction (from the most recent non-speech region) no longer matches (and is hence unable to subtract out) the time-varying color noise spectra in the noisy speech in subsequent speech regions. In contrast, since the Wiener filter uses the clean speech estimate in addition to the noise - estimate, Wiener filtering is able to provide an improved enhancement by virtue of having a good approximation of the underlying speech spectrum during the noisy speech period even in such conditions when the noise-estimate is inadequate to correctly represent the current noise spectra in the speech regions.

These differences are brought out in the following experimental results **for white noise**. Fig. 4 shows the closed-set speaker-identification accuracy using the one-pass DP algorithm with 1 and 5 templates for test data SNRs of 0 dB, 5 dB and 10 dB. It can be seen that, while the noisy speech (Noisy) has a very poor per-

formance, spectral subtraction (SS) provides only a marginal improvement. However, HWF has a significantly high performance, clearly validating the effectiveness of the proposed algorithm for speaker-identification for all the SNRs considered here. The performance improvement (about 10%) from using 1 to 5 templates in the one-pass DP algorithm [6] can also be noted.

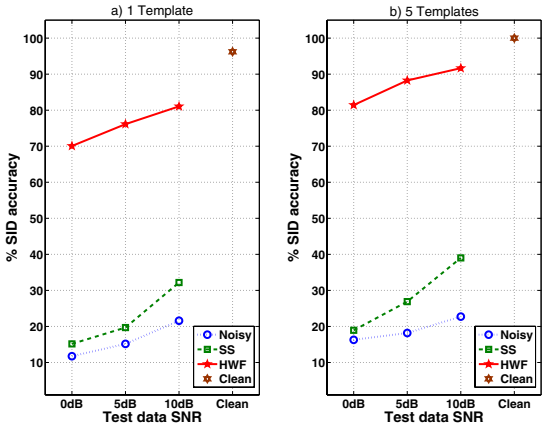

Figure 4: *Closed-set speaker-identification accuracy (%)*

Fig. 5 shows the speaker-verification performance using the DET (Detection error trade-off) curve. The one-pass DP algorithm is used with 5 templates for a test data SNR of 0 dB. It can be noted that, as in speaker-identification, the performance of noisy speech is very poor and spectral subtraction does not improve this. On the contrary, HWF offers an excellent improvement with a highly lowered EER (Equal-error-rate) where the probability of false acceptance (p_{fa}) equals the probability of false rejection (p_{fr}). Table 1 shows the EER points (p_{fa}, p_{fr}) for various SNRs (0 dB, 5 dB and 10 dB) for all the three cases – Noisy speech (NOISY), spectral subtraction (SS) and the proposed HWF algorithm (HWF) using 5 templates in the one-pass DP algorithm (as in Fig. 5). Here again, it can be noted that HWF offers the best performance improvement, while SS performs as poorly as the NOISY case itself.

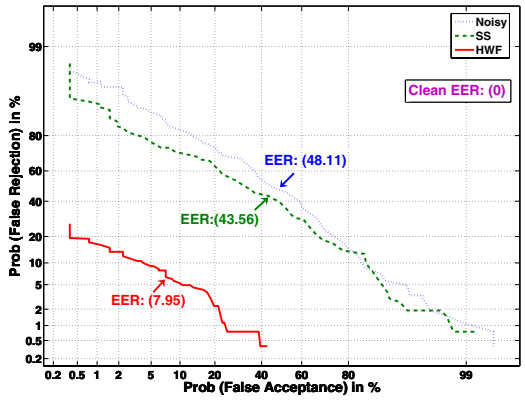

Figure 5: *Speaker-verification DET plots for 0 dB test data SNR*

In order to show the effectiveness of HWF on **non-stationary noise** as discussed in the earlier part of this section, we evaluated the closed-set speaker-identification performance of the algorithm on three types of noises, namely, factory noise, chopper noise, and babble noise for a test data SNR of 0 dB for the same set of speakers in TIDIGITS as above (and with 5 templates in the one-pass DP algorithm). Table 2 shows the speaker-identification accuracy for the noisy speech (NOISY), spectral subtraction (SS) and the proposed HWF algorithm. While the performance of noisy speech is poor, spectral subtraction achieves only modest relative improvements over the noisy case. However, HWF has an excellent performance offering a large improvement over the noisy and SS cases.

Table 1: *Speaker-verification EER points (p_{fa}, p_{fr}) for test SNRs (0 dB, 5 dB, 10 dB); clean EER=(0,0)*

Test SNR	NOISY		SS		**HWF**	
	p_{fa}	p_{fr}	p_{fa}	p_{fr}	p_{fa}	p_{fr}
0 dB	48.11	47.35	43.56	43.18	**7.95**	**7.20**
5 dB	46.59	44.70	36.74	34.85	**3.79**	**4.17**
10 dB	42.05	42.80	28.79	31.82	**2.27**	**1.89**

Table 2: *Speaker-identification accuracy (%) for 3 non-stationary noises for test SNR of 0 dB; clean accuracy = 100%*

Noise type	NOISY	SS	**HWF**
Factory	17.04	28.41	**95.45**
Chopper	29.54	48.86	**88.63**
Babble	52.27	79.54	**96.50**

5. Conclusions

We have proposed a highly noise robust text - dependent speaker - recognition algorithm based on hypothesized Wiener filtering (HWF). The proposed algorithm exploits the availability of the clean reference templates of the words of the password text (supposed to be the text of the input noisy speech) in text-dependent mode of operation. The proposed HWF algorithm is set within the one-pass DP framework and is evaluated for both speaker-identification and speaker-verification using the TIDIGITS database. The proposed algorithm has very high recognition accuracies for both additive white-noise and non-stationary color noise.

6. References

[1] T. F. Quatieri. *Discrete-time speech signal processing – Principles and practice*. Pearson Education, 2002.

[2] A. D. Berstein and I. D. Shallom. An hypothesized wiener filtering approach to noisy speech recognition. In *Proc. ICASSP'91*, pages 913–916, 1991.

[3] V. L. Beattie and S. J. Young. Noisy speech recognition using hidden Markov model state-based filtering. In *Proc. ICASSP'91*, pages 917–920, 1991.

[4] S. V. Vaseghi and B. P. Milner. Noisy speech recognition based on HMMs,Wiener filters, and re-evaluation of most likely candidates. In *Proc. ICASSP'93*, pages II–103–II–106, 1993.

[5] V. Ramasubramanian and Amitav Das. Text-dependent speaker-recognition – A survey and state-of-the-art. Tutorial at ICASSP-2006, Toulose, France, May 2006.

[6] V. Ramasubramanian, Amitav Das, and V. Praveen Kumar. Text-dependent speaker-recognition using one-pass dynamic programming algorithm. In *Proc. ICASSP'06*, pages I-901–I-904, Toulose, France, May 2006.

[7] V. Ramasubramanian et al. Text-dependent speaker-recognition systems based on one-pass dynamic programming algorithm. In *Proc. IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, Puerto Rico, June 2006.

[8] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustics, Speech and Signal Proc.*, 27(2):113–120, Apr 1979.