Proceedings of the Fourth IASTED International Conference
Signal Processing, Pattern Recognition, and Applications
February 14–16, 2007, Innsbruck, Austria
ISBN Hardcopy: 978-0-88986-646-1 / CD: 978-0-88986-647-8

ROBUST SPEAKER VERIFICATION IN AIR TRAFFIC CONTROL USING IMPROVED VOICE ACTIVITY DETECTION

Michael Neffe and Tuan Van Pham
Signal Processing & Speech Communication Lab.
Graz University of Technology, Austria
email: michael.neffe@TUGraz.at and v.t.pham@TUGraz.at

Franz Pernkopf and Gernot Kubin
Signal Processing & Speech Communication Lab.
Graz University of Technology, Austria
email: pernkopf@TUGraz.at and g.kubin@ieee.org

ABSTRACT
In this paper, a robust speaker verification system using improved voice activity detection has been designed for increasing safety of air traffic control. In addition to the usage of the aircraft identification tag to assign speaker turns on the shared communication channel to aircrafts, speaker verification is investigated as an add-on attribute to improve security level effectively for the air traffic control. The front-end processing unit is optimized to deal with small bandwidth restrictions and very short speaker turns. Two adaptive voice activity detection methods based on energy and wavelet parameters are developed and used as pre-processing in front-end unit. The verification task is accomplished by training background models and speaker dependent models. To enhance the robustness of the verification system, a cross verification unit is further applied. The designed system is tested with SPEECHDAT-AT and WSJ0 database to demonstrate its superior performance.

KEY WORDS
Speaker verification, Gaussian mixture model - universal background model, voice activity detection, quantile filtering, wavelet transform.

1. Introduction

Speaker verification (SV) is an approach to identify a speaker from captured speech signal. This technique shows potential possibilities of different voice-controlled applications such as dialog system, information services and security control. In our research, SV is applied to increase security level of Air Traffic Control (ATC). There is a steady demand for increasing the security level in ATC voice communication between controller and pilots. In 2003 the Eurocontrol Experimental Centre (EEC) [1] proposed the Aircraft Identification Tag (AIT) which is based on a watermarking technique to identify the originating aircraft of the transmitting voice source [2]. A further improvement of the security level was proposed in [3] by using a SV system based on the AIT information. Here the pilot's voice is automatically enrolled when the pilot registers the first time to a control sector. At any later occurrence of the same AIT the new received voice message is verified against the existing speaker model. The main challenge for the ATC-oriented SV system is degradation of the transmitted noisy signal caused by the fading channel [4]. The bandwidth limitation of the transceiver equipment is also a critical point. ATC uses a bandwidth of only 2200 Hz in the range of $300 - 2500$ Hz [3] for speech transmission. The short duration of pilots' speaker turn speech is another challenge. Hering et al. [2] have shown that one speaker turn is only five seconds on average in length during pilots and controller communication.

To solve the above demands the text-independent Gaussian mixture model - universal background model (GMM-UBM) approach is used for speaker verification with a front-end processing unit specially adapted to the needs of ATC. An improved voice activity detection (VAD) is embedded as pre-processing in front-end unit to extract exact speech segments. VAD has a main impact on SV performance under noisy conditions which helps to raise the security level as studied in [3]. Voice activity detection (VAD) is a most crucial topic in speech processing and its application. Many VAD techniques have been proposed using single-domain features such as short-term energy levels, zero crossing rate, autocorrelation coeficients, glottal closure indices [5]. Other methods are based on multi-domain features extracted from the short-time Fourier transform (STFT) of the input speech frames such as mel frequency cepstral coefficients [6]. Recently, the Wavelet transform (WT) which provides a flexible rectangular tiling of the time-frequency plane is applied for phonetic classification in noisy environments [7].

In this paper, an improved VAD is developed from a wavelet-based phonetic classification [7] with a novel adaptive quantile filtering method. The wavelet coefficients obtained by the WT on every windowed overlapping speech frame are used to extract a frame-based delta feature. To estimate noise threshold, a quantile filtering method which is proposed in [7] is further improved by an adaptive estimation of the quantile factor. For smoothing of fluctuations resulting from strong non-stationary noise in the VAD outputs, a hangover scheme [3] has been applied. The speech segments derived at the VAD outputs are used to extract linear-frequency cepstral coefficients (LFCCs). After that, gender-dependent universal background models (UBMs) are trained. Speaker dependent models (SDMs) are adapted from the UBM for each speaker before the verification task is performed.

The paper is structured as follows: An improved VAD based on the WT and adaptive quantile filtering is presented in section 2. Then the design of the SV system is explained in section 3. Experiments and discussion can be found in section 4. The final section presents a conclusion and future research.

2. Improved Voice Activity Detection

2.1 Multi-resolution Analysis and Feature Extraction

Based on the multi-resolution capabilities of the WT, any discrete-time signal $x[k]$ can be decomposed into the sum of an approximation plus L details at L^{th} scale as:

$$\begin{aligned} x[k] = & \sum_{n=1}^{N_s} X^{(L)}[2n] \cdot g_0^{(L)}[k - 2^L n] + \\ & \sum_{m=1}^{L} \sum_{n=1}^{N_s} X^{(m)}[2n+1] \cdot g_1^{(m)}[k - 2^m n], \end{aligned} \tag{1}$$

where $X^{(L)}[2n]$ and $X^{(m)}[2n+1]$ are the approximation coefficients (low-frequency part) and the detail coefficients (high-frequency part), respectively. They are defined as:

$$\begin{aligned} X^{(L)}[2n] &= \left\langle h_0^{(L)}[2^L n - l], x[l] \right\rangle, \\ X^{(m)}[2n+1] &= \left\langle h_1^{(m)}[2^m n - l], x[l] \right\rangle, \end{aligned} \tag{2}$$

where $g_0^{(m)}[k]$ is an equivalent filter obtained through m stages of lowpass synthesis filters $g_0[k]$, each preceded by an upsampler by 2. $h_0^{(m)}[k]$ is an equavalent lowpass analysis filter. We call $W_{m,i}(n)$ the sequence of all wavelet coefficients (i.e, $X^{(L)}[2n]$ and $X^{(m)}[2n+1]$) which are derived by the WT at the m^{th} scale of the i^{th} frame, n is the coefficient index, N_s is the number of wavelet coefficients in each subband, $m, n, k, i \in \mathbb{Z}$.

As observed by Pham et. al [7], there are different wavelet power distributions for different phonetic classes of speech signals. A relatively uniform power distribution occurs for the non-speech frames. However, the power of the voiced frames is mostly contained in the approximation subbands and much less in the detail subbands, and vice versa for the unvoiced frames. These significant power differences between approximations and details are used to detect the speech frames. From the statistical properties of speech sounds, we see that the spectrogram power in the range $[0-1]$kHz is very high for voiced frames in comparison with unvoiced frames. Dealing with the air traffic voice signal having a bandwidth restricted to the range of $[0.3-2.5]$ kHz, we choose a decomposition scale $m=2$ to consider the relation between a low-frequency band of $[0.3-1.1]$kHz and the remaining higher-frequency band. A delta parameter $D(i)$ which is the power difference between approximation and detail subbands is calculated for each speech frame as follows:

$$D(i) = \frac{1}{N_a} \sum_{n=1}^{N_a} W_{m,i}^2(n) - \frac{1}{N - N_a} \sum_{n=N_a+1}^{N} W_{m,i}^2(n). \tag{3}$$

where $N_a = \dfrac{N}{2^m}$ and $N - N_a$ are the length of the approximation subband and detail subbands, respectively. N is the number of samples in one speech frame.

2.2 Robust Feature

Due to the fading channel of air traffic communication, the transmitted voice signal is degraded significantly. Thus, the hyperbolic tangent sigmoidal function is applied on $D(i)$ in order to amplify small values of the delta feature $D(i)$ resulting from weak speech frames. This operation also compresses very high values of the $D(i)$ resulting from high quality speech frames to balance the impact of the large range of values of $D(i)$ during processing.

In addition to voice signal distortion, we observe that there are high fluctuations of the feature values $D(i)$ during non-speech segments. This results from strong non-stationary noise of the transmitting channel. To make the VAD robust against noise, the processed delta features are further smoothed by median filtering of length five frames:

$$D_s(i) = \text{medfilt}\left(\frac{1 - e^{-2D(i)}}{1 + e^{-2D(i)}} \right). \tag{4}$$

2.3 Adaptive Noise Threshold

In order to achieve accurate speech/non-speech detection, the extracted feature values will be compared with an estimated threshold. A statistical quantile filtering method proposed in [7] is further improved to have a better estimate of the threshold relating to the noise level. We observe that the smoothed feature values $D_s(i)$ stay at the noise level over a significant part of buffers ten seconds in length. The estimation is implemented in two steps:

- Sorts $D_s(i)$ in ascending order over a buffer b of N_f frames to get $D_s(i'),\ i' = [1 \dots N_f]$.
- Determines an adaptive threshold $T_q(b)$ by taking the q^{th} quantile:

$$T_q(b) = D_s(i')|_{i' = \lfloor qN_f \rfloor} \tag{5}$$

With this method, the threshold is updated for every captured speech buffer and is adaptive to non-stationary noise which is common in ATC. As studied in [7], the quantile factor $q = 0.3$, which had been selected experimentally from the range of possible values $q = [0.0 \dots 1.0]$, provides a good estimate of the noise threshold. However, this constant quantile factor for every buffer introduces a limitation of the method. From the temporal characteristics of the input utterances, we observe that the ratio between the number of speech frames and the number of non-speech frames varies for different buffers. An adaptive quantile factor $q(b)$ is proposed to achieve a better estimate of the noise threshold. The method is based on a comparison of feature difference between every five consecutively sorted frames and a pre-determined level $\varepsilon = 10^{-3}$. The procedure is done from the beginning of the buffer and is stopped when the difference is larger than the level. Then the quantile factor $q(b)$ is selected as:

$$q(b) = i', \qquad \text{if} \quad D_s(i') - D_s(i'-4) > \varepsilon \tag{6}$$

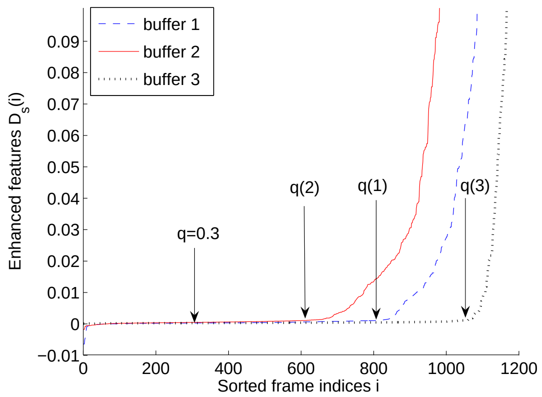

Figure 1: Adaptive quantile factors for different buffers.

As shown in Fig. 1, three different quantile factors $q(b)$ are estimated more accurately than the constant quantile factor for three different buffers. To make the speech/non-speech decisions, the smoothed delta parameter $D_s(i)$ of each input speech frame is calculated and compared with the estimated threshold. They are labeled as speech frames if the absolute values of $D_s(i)$ are larger than the threshold $T_q(b)$, and as non-speech frames otherwise:

$$\text{VAD}(i) = \begin{cases} \text{Speech}, & \text{if } |D_s(i)| > T_q(b) \\ \text{NonSpeech}, & \text{otherwise} \end{cases} . \tag{7}$$

Finally, the output sequence $\text{VAD}(i)$ is smoothed by applying a 100ms/ 200ms hangover scheme in [3] to bridge short voice activity regions, preserving only candidates with a minimal duration of 100 ms, and being not more apart than 200 ms from each other. This excludes talk-spurts shorter than 100 ms and relabels pauses smaller than 200 ms. The impact of the bridging rule is considered during experiments. A block scheme of the proposed voice activity detector is presented in the following figure:

3. Speaker Verification Design

Based on the detected VAD segments, features are extracted and normalized. The design of the SV system consists of four phases as shown in Figure 3. In phase 1, gender dependent UBMs are trained. These models are used in phase 2 for speaker dependent modeling using gender information from gender recognition. Retraining of a speaker model is performed in phase 3 and finally in phase 4 the verification task is carried out.

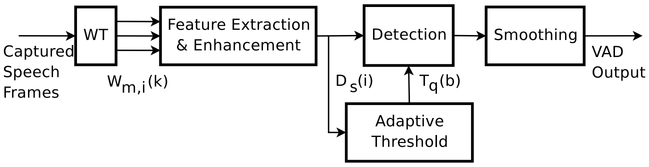

Figure 2: Block scheme of the improved VAD.

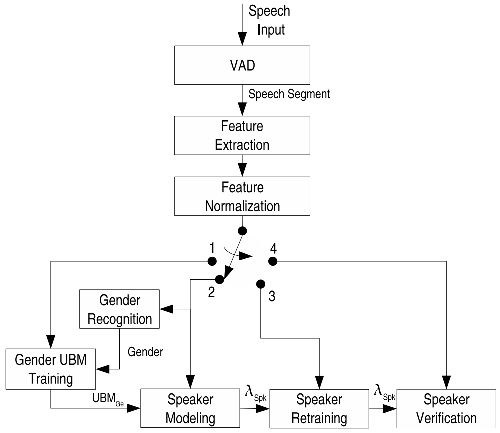

Figure 3: Illustration of designing a SV system in 4 phases.

3.1 Front-end Processing

A robust VAD is crucial in the SV system in order to extract suitable speaker dependent data. Non-speech data which is contaminated by noise of the transmission channel may drive model training into incorrect convergence. This leads to an unreliable SV system.

Feature Extraction: For each speech segment detected by the VAD method features are extracted separately as shown in Fig. 3. This is necessary to avoid artificial discontinuities when concatenating feature values of speech frames. 14 cepstral coefficients are extracted using a linear frequency, triangular shaped filter bank with 23 channels between 300 Hz and 2500 Hz for each frame. As proposed in [3] a frame length of 25 ms and a frame rate of 5 ms achieves good results. Finally the whole feature set comprises these LFCCs calculated in *dB* and the polynomial approximation of its first and second derivatives [8]. Altogether 42 features per frame are used.

Feature Normalization: In order to reduce the impact of channel dependent distortions histogram equalization (HEQ) [9] has been used as feature normalization method as shown in Fig. 3. HEQ is known to normalize not only the first and the second moment but also higher-order ones. The HEQ method maps an input cumulative histogram distribution onto a Gaussian target distribution. This distribution is calculated by sorting the input feature distribution into 50 bins. This number has been selected to be small to get sufficient statistical reliability.

3.2 SV Classification

Here we use the GMM-UBM approach first introduced by [10]. In contrast to other GMM-UBM SV systems [8] we decided to train gender dependent UBMs which are finally not merged to one global UBM. For training the UBM, the basic model has been initialized randomly and then trained in a consecutive manner by the speech data using maximum a posteriori (MAP) adaptation. For retraining of the model to obtain the final gender dependent UBM, we used three EM - steps and a weighting factor directly proportional to the ratio of the total speech length used so far for training and the new utterance length, the model is going to be retrained too. This is done in phase 1 as shown in Fig. 3. To form a SDM, first the gender is determined according to the log-likelihood of the gender-models as:

$$Ge = \arg\max \left[L(X, \lambda_{UBM_m}), L(X, \lambda_{UBM_f}) \right], \quad (8)$$

where $L(X, \lambda)$ is the log-likelihood of the model λ given the data X, f and m depict the female and male UBMs, respectively. The corresponding gender dependent UBM is used to adapt a SDM in phase 2. For speaker adaptation three EM - steps and a weighting factor of 0.6 for the adapted model and correspondingly 0.4 for the UBM is used to merge these models to the final SDM. In phase 3 further adaptation of the SDM with new data is done by retraining the model as described for the UBM retraining.

The score $S(X)$ which is used for verification in phase 4 is calculated by comparing the hypothesized speaker namely the speaker model λ_{Spk} with its anti-hypothesis the UBM $\lambda_{UBM_{Ge}}$:

$$S(X) = \log L(X|\lambda_{Spk}) - \log L(X|\lambda_{UBM_{Ge}}). \quad (9)$$

3.3 Cross Verification

To meet the high security expectations in ATC voice communication a cross verification unit can be applied as add-on. If an utterance is shorter than a predefined minimum length (i.e., 8 seconds) and the score is not confident enough (positive or negative) the system waits for another utterance and conducts a cross verification as proposed in [3]. Therefore, let X_1 and X_2 be the sequence of feature vectors of the first and second utterance to be investigated and λ_1 and λ_2 their adapted speaker models, respectively. If $S_{\lambda_2}(X_1) \cap S_{\lambda_1}(X_2) > t$, i.e., both scores are above a threshold *t* and are verified to be from the same gender as defined in Eqn. 8, then it is assumed that both utterances are from the same person and thus are concatenated and used for verification. Figure 4 shows the region of insufficient confidence in the score distribution histogram. Intruders and true speakers are illustrated separately. The region of low confidence which is shown as dashed box in the figure has been set to -1.8 ± 0.2 using the energy-based VAD.

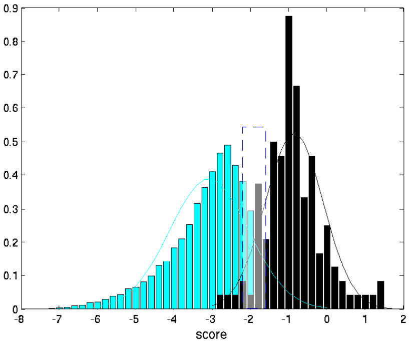

Figure 4: Histogram and fitted Gauss curves for the score distributions of imposters (left) and true speakers (right). The dashed rectangle illustrates the score region of low confidence.

4. Experiments and Discussion

The fixed telephone SPEECHDAT-AT database [11] and the WSJ0 database [12] are used in our experiments. Dialect regions and speaker ages were assumed to be selected randomly. In order to simulate the conditions of ATC, all files were band-pass filtered to a bandwidth from 300 Hz to 2500 Hz and down-sampled to a sampling frequency of 6 kHz. To match ATC conditions the databases were cut artificially in utterances of 5 seconds which corresponds to a typical talk spurt length in ATC. For training/retraining a SDM, 3 such segments are used in a row. For the experiment a total of 200 speakers comprising 100 females and 100 males were randomly chosen from the SPEECHDAT-AT database. Gender-dependent UBMs were trained with 38 Gaussian components using two minutes of speech material for each of 50 female/male speakers. Out of the remaining 100 speakers 20 were marked as reference speakers. Both, for the remaining 99 speakers, known as imposters as well as for the reference speakers, 6 utterances were used for verification. So each reference speaker was compared to 600 utterances, yielding a total of 12000 test utterances for 20 reference speaker models all together.

For the tests conducted on the WSJ0 database the CD 11_2_1 comprising 23 female and 28 male speakers was used to train the gender dependent UBMs. Since in this database each speaker produces the same utterances, 100 seconds of speech were randomly selected from each speaker and used for training. For testing CD 11_1_1 with 45 speakers divided into 26 female and 19 male ones were taken. Here again the speech files for the reference speaker as well as for the claimants were selected randomly but have been the same for all different VAD experiments.

Speech material used for training/retraining the reference speaker was labeled and hence excluded from verification - 24 were labeled as reference speakers, 12 female and 12 male each. Both, for the remaining 44 speakers as well as for the reference speaker, 12 utterances were used for verification. So each reference speaker was compared to 540 utterances which yields a total number of 12960 test utterances for 24 reference speakers.

Results: To measure SV performance we use the Detection Error Tradeoff (DET) curve and as special point in this curve the equal error rate (EER). The score distribution of all speaker utterances are illustrated in Fig. 5. The reference speaker utterances used to test the reference model are highlighted in black for all reference models along the x-axis. Imposters are depicted in grey for certain score values on the y-axis. The z-axes shows the score for different utterances of imposters and claimants respectively.

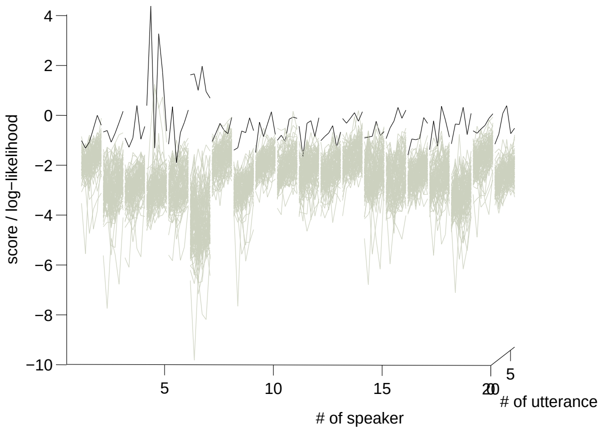

Figure 5: Score distribution for an experiment on SPEECHDAT-AT for all 20 reference speakers (black) and all 100 claimants (grey) separately using the WT based VAD.

EER [%]	NoVad	EVad wo/w	WaVad wo/w
SPEECHDAT-AT	25.12	11.7 / 6.52	9 / 4.75
WSJ0	10.15	- / 10.37	- / 10

Table 1: EER results derived from both databases for different VADs without (wo) and with (w) applying hangover scheme.

The energy-based VAD in [3] is used to compare with the proposed WT-based VAD in term of SV performance. As reported in Table 1, for the SPEECHDAT-AT database which consists of noisy fixed line telephone recordings, the usage of both VAD methods improves SV performance significantly compared to the case without using VAD. However, for the almost noise-free WSJ0 database, the obtained results are almost similar. This shows a positive effect of VAD in removing noise-dominated non-speech segments which may lead to an unreliable trained SV system. With the more accurate WT-based VAD than the energy-based VAD, the EER is reduced from 11.7% to 9 % without smoothing, and from 6.52% to 4.75% with smoothing as illustrated in Fig. 6. Thus, by using the proposed WT-based VAD, we gain 23% and 27% relative improvement compared to the energy-based VAD in both cases. In addition, from the observed results, we discovered that not only an accurate detection of speech frames but also a smoothing to bridge short pauses between speech frames help to improve the SV performance.

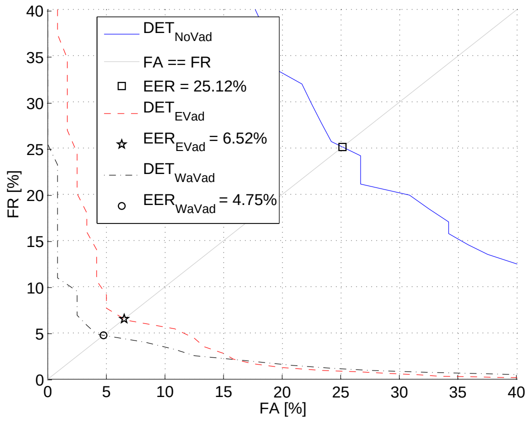

Figure 6: DET-curve, as plot of false acceptance (FA) rate versus false rejection (FR) rate and EER point for the SV system without VAD (NoVad), energy-based (EVad) and WT-based VADs (WaVad).

The impact of the proposed cross verification unit has been studied on SPEECHDAT-AT database for the energy-based VAD only. As shown in Fig. 7 we can report a reduction of the EER from 6.52 % to 6.12 %.

To assess the impact of environmental mismatch between training and test conditions, a cross testing has been performed using SPEECHDAT-AT database for training UBMs but WSJ0 database for testing and vice versa. The WT-based VAD is employed for these experiments. In the former condition, the EER is 11.8 % which is worse than above results because the models were trained by noisy speech and tested with clean speech. In the later condition, the slight improvement of EER to 11 % may result from the effect of VAD in reducing of noisy non-speech segments in testing phase. In both conditions, using VAD can not solve the mismatch between training and testing phases.

5. Conclusion

In this paper, a voice activity detection is built in front-end processing stage to improve speaker verification performance. Our system was specially designed for the needs of

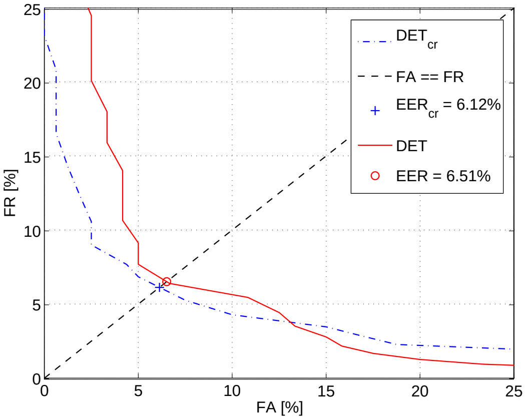

Figure 7: DET curves with EER point (plus sign and circle) for both the normal system and the system with cross-verification. Subscript *cr* denotes the usage of the cross-verification system.

air traffic control with respect to bandwidth restriction and talk spurt length. Systematic tests have been performed using this system without VAD, with the energy-based VAD and the WT-based VAD . The last method which is based on enhanced wavelet feature and adaptive noise threshold provides best performance. The hangover scheme which smooths the VAD output contributes to the EER reduction. Significant EER reduction is achieved under harsh environments when applying the WT-based VAD, and hence it proves indispensable for further security enhancements in air traffic control voice communication.

For future work, parameters of the WT-based VAD will be fine tuned to maximize EER performance. Relationship between VAD performance and speaker verification performance should be considered. An adaptive estimate of the low confidence region for the cross-verification unit will be investigated. Moreover, we want to examine results on the whole SPEECHDAT-AT database and try UBM environment adaptation for compensating mismatch conditions in training and testing. As another approach to reduce the impact of environmental mismatch, noise reduction can be applied as pre-processing step as studied for robust automatic speech recognition in [13].

Acknowledgment

The authors wish to thank Horst Hering from EEC for his assistance throughout this work.

References

[1] “Eurocontrol Experimental Centre-EEC,” Brétigny-sur-Orge, France.

[2] H. Hering, M. Hagmüller, and G. Kubin, “Safety and security increase for air traffic management through unnoticeable watermark aircraft identification TAG transmitted with the VHF voice communication,” in *Proc. of the* 22^{nd} *IEEE Dig. Avionics Sys. Conf., DASC '03.*, 2003, vol. 1, pp. 4.E.2–41–10 vol.1.

[3] M. Neffe, T. V. Pham, H. Hering, and G. Kubin, *Speaker Classification*, chapter Speaker Segmentation for Air Traffic Control, Springer: Lecture Notes in Artificial Intelligence, accepted for publication.

[4] Hofbauer K., Hering H., and Kubin G.:, “A measurement system and the TUG-EEC-Channels database for the aeronautical voice radio,” in *IEEE Vehicular Technology Conference*, Montreal, Canada, 2006.

[5] D. G. Childers, *Speech processing and synthesis toolboxes*, John Wiley & Sons, USA, 2000.

[6] Z. Xiong and T. Huang, “Boosting speech/non-speech classification using averaged mel-frequency cepstrum,” in *Proc. Pacific-Rim Conference on Multimedia*, 2002.

[7] T. V. Pham and G. Kubin, “Low-complexity and efficient classification of voiced/unvoiced/silence for noisy environments,” in *Proc. Interspeech*, 2006.

[8] F. Bimbot and et al., “A tutorial on text-independent speaker verification,” in *EURASIP Journal on Applied Signal Processing*, 2003, number 4, pp. 430–451, EURASIP.

[9] M. Skosan and D. Mashao, “Modified segmental histogram equalization for robust speaker verification,” *Pattern Recognition Letters*, vol. 27, no. 5, pp. 479–486, Apr. 2006.

[10] F. Bimbot and et al., “Speaker verification using adapted gaussian mixture models,” in *Digital Signal Processing*, 2000, number 10, pp. 19–41.

[11] M. Baum, G. Erbach, and G. Kubin, “SPEECHDAT-AT: A telephone speech database for Austrian German,” Proc. LREC Workshop Very Large Telephone Databases, 2000.

[12] J. Garofalo, D. Graff, D. Paul, and D. Pallett, “Continous Speech Recognition (CSR-I) Wall Street Journal (WSJ0) news, complete,” Linguistic Data Consortium, Philadelphia, 1993.

[13] E. Rank, T. V. Pham, and G. Kubin, “Noise suppression based onwavelet packet decomposition and quantile noise estimation for robust automatic speech recognition,” in *Proc. ICASSP*, 2006, vol. 1, pp. 477–480.