Robust Voice Activity Detection For Narrow-Bandwidth Speaker Verification Under Adverse Environments

Tuan Van Pham, Michael Neffe, Gernot Kubin

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Austria
v.t.pham@TUGraz.at, michael.neffe@TUGraz.at, g.kubin@ieee.org

10.21437/Interspeech.2007-169

Abstract

We describe a voice activity detection algorithm which leads to significant improvement of a narrow-bandwidth speaker verification system under harsh environments. This algorithm is based on a time-scale feature which is extracted from wavelet subbands. A statistical quantile filtering technique is proposed to estimate an adaptive noise threshold. A hang-over scheme is then applied to bridge short pauses between speech frames. This optimized voice activity detector is embedded in the front-end unit of the narrow-bandwidth speaker verification system. The proposed algorithm is evaluated by objective tests on band-pass filtered utterances from the TIMIT database which was artificially corrupted by different additive noise types. Furthermore, it is tested with band-pass filtered SPEECHDAT-AT and WSJ0 databases in terms of speaker verification rates. This algorithm shows its superiority in performance due to the robust time-scale feature and the adaptive threshold.

Index Terms: robust voice activity detection, speaker verification, statistical quantile filtering, time-scale feature.

1. Introduction

Reliable voice activity detection (VAD) is a most crucial topic among many approaches to robust speech processing. Very common features are used for this task such as short-term energy, zero-crossing rate, autocorrelation coefficients [1]. These features may be affected by complex and strong noise. Some methods have been proposed to deal with harsh conditions such as the features extracted from the frequency domain, mel frequency cepstrum coefficients [2]. The application of the time-scale features which are derived from the discrete wavelet transform (DWT) has been investigated recently. By employing the auto-correlation function and the Teager energy operator (TEO) calculated from wavelet subbands, a voice detector has been designed in [3]. Another method using subband order-statistics filters (SOSF) [4] can improve VAD performance in non-stationary noise conditions. VAD is used to estimate the noise level for many noise reduction methods [5]. In automatic speech recognition, there is a need for speech/non-speech detection to improve the recognition rate [6]. Application of VAD in speaker verification (SV) systems was introduced in [7] to achieve reliable model estimation. Furthermore, the optimization of VAD to ensure robustness of the narrow-bandwidth SV system operating in adverse conditions such as air traffic control (ATC) [8] is a challenging task.

In this paper, we develop a robust VAD from a phonetic classifier which was proposed in [9] for improving performance of the narrow-bandwidth SV system. The proposed VAD is applied in the front-end unit to extract only voice segments for the purposes of training the system and the verification task. Firstly, a time-scale feature is calculated from the TEO of the wavelet coefficients which were derived by applying the DWT at the 1 scale on every speech frame. Secondly, a sigmoid function and median filter are applied on the extracted feature to make it robust against noise. A statistical quantile filtering (SQF) method using adaptive quantile factors is employed to estimate a threshold relating to the noise level accurately in case of non-stationary noise. Finally, a hang-over scheme is applied to smooth out fluctuations. The obtained voice activity segments are used to extract linear-frequency cepstral coefficients (LFCCs). Gender-dependent universal background models (UBMs) are trained after that. By adapting the UBM for each speaker, we derive speaker dependent models (SDMs) for the verification task.

The paper starts with a description of the specific SV system in section 2. Then a robust VAD method which is based on the time-scale feature and statistical quantile filtering is presented in section 3. In section 4, the performance of the VAD and the SV system is evaluated and discussed. The final section presents a conclusion and future research.

2. Speaker Verification System

Speaker verification (SV) serves to verify a claimed speaker identity. In this study, a SV system is applied as an enhancement to increase safety in ATC. The main challenges for the ATC-oriented SV system are the noise corrupting the transmitted signal and the limited frequency range of only $300 - 2500$ Hz used for speech transmission [8]. Due to the lack of an ATC speech database, we mainly deal with the limitations of narrow-bandwidth and partially noisy conditions by simulating these conditions on other databases. The design of the SV system consists of four phases as shown in Figure 1. In phase 1, gender-dependent UBMs are trained offline. These trained models are applied in phase 2 for speaker dependent modeling using recognized gender information. Retraining of a speaker model is performed in phase 3, and finally, in phase 4 the verification task is carried out. A detailed description can be found in [8].

This research was carried out in the context of COAST-ROBUST, a joint project of Graz University of Technology, Philips Speech Recognition Systems, and Sail Labs Technology. We gratefully acknowledge funding by the Austrian KNet Program, ZID Zentrum fuer Innovation und Technology, Vienna, the Steirische Wirtschaftsfoerderungsgesellschaft mbh, and the Land Steiermark. For further information, see http://www.coast.at. Additional support in the ATC domain was provided by the Eurocontrol Experimental Centre, Brétigny, France

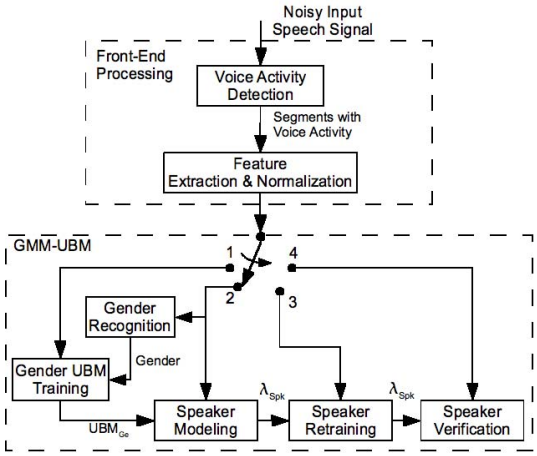

Figure 1: VAD as pre-processing stage in the SV system.

2.1. Feature Extraction and Normalization

Every windowed frame of 25 ms length with 5 ms frame rate is taken from the input singal to extract features. These frames are first Fourier transformed and further processed by a triangular shaped, linear filterbank with 23 channels between 300 and 2500 Hz. Finally 14 cepstral coefficients are calculated by applying the inverse discrete Cosine transform. The final feature set consists of these LFCCs and its first and second differences, resulting in 42 features altogether. To enhance robustness against mismatches between training and test conditions due to changing environments, channels, or handset types Histogram equalization (HEQ) has been carried out. The HEQ method maps an input cumulative histogram distribution in our system onto a Gaussian target distribution. This distribution is calculated by sorting the feature distributions into a small number of only 50 bins in order to get sufficient statistical reliability of the data in each bin.

2.2. Speaker Verification Classification

For classification a similarity-based method, namely the Gaussian mixture model-UBM approach first introduced by [10], has been used. For this system we decided to train gender dependent UBMs which are finally not merged to one global UBM. The UBM training in phase 1 has been done in a consecutive manner by the speech data starting with a randomly initialized model. For the retraining of the model to yield the final gender dependent UBM, we used three EM - steps and a weighting factor dependent on the speech data length used. Based on the enrollment data, one of the gender-dependent background models is selected as a seed model by the gender detector:

$$gender = \max_{\text{Ge} \in \text{ female, male}} L(X|\lambda_{\text{UBM}}^{\text{Ge}}) \quad (1)$$

where $L(X|\lambda)$ is the log-likelihood of the gender dependent UBMs $\lambda_{\text{UBM}}^{Ge}$ given the data X. The corresponding gender dependent UBM is used to adapt a SDM as shown for phase 2. In phase 3 further adaptation of the SDM with new data acquired during the online use of the system, is done by retraining the existing model. Experimental tests resulted in a slightly better performance for updating not only the means and priors but also the diagonal variances in phases 2 and 3. The score $S(X)$ which is used for verification in phase 4 is calculated by comparing the hypothesized speaker model λ_{Spk} with its anti-hypothesis $\lambda_{\text{UBM}}^{\text{Ge}}$:

$$S(X) = L(X|\lambda_{\text{Spk}}) - L(X|\lambda_{\text{UBM}}^{\text{Ge}}) \quad (2)$$

As depicted in Fig. 1, VAD is employed in the front-end unit as an important step to detect speech segments only, which are used to extract suitable speaker dependent data. Non-speech data contaminated by noise of the transmission channel may drive the model training process into incorrect convergence, thereby leading to an unreliable SV system. The design of the VAD is presented in the following section.

3. Robust Voice Activity Detection

Based on a phonetic classifier for voiced, unvoiced and silence classes which was built in [9], we develope a robust voice activity detector optimized for the narrow-bandwidth SV system shown in the block diagram in Figure 2.

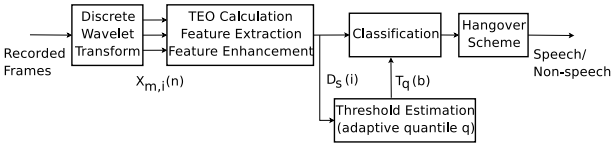

Figure 2: Block scheme of the robust VAD.

3.1. Reliable Time-Scale Feature Extraction

To detect voice activity segments accurately in adverse environments, a time-scale feature is extracted in the wavelet domain. This feature extraction is based on a useful observation reported by Pham et. al in [9]. It is observed that a relatively uniform power distribution between the approximation and the detail subbands occurs for the non-speech frames. However, there are significant power differences between approximations and details of speech frames consisting of voiced and unvoiced frames. From the statistical properties of speech sounds, we observed that the power in the range of $[0-1]$kHz is very high for voiced frames in comparison with unvoiced frames. Dealing with narrow-bandwidth speech signal from $[0.3-2.5]$ kHz as common in ATC, we choose a decomposition scale $m=1$ to consider the relation between a low-frequency (approximation) wavelet subband of $[0.3-1.1]$kHz and the remaining higher-frequency (detail) wavelet subband. By applying the DWT at the m^{th} scale of the i^{th} windowed speech frame, we derive the sequence of wavelet coefficients $X_{m,i}(n)$. It consists of approximation coefficients $a_{m,n}$ and detail coefficients $d_{m,n}$ calculated by the following discrete convolutions:

$$a_{m,n} = \sum_p a_{m-1,p}\, h(p-2n) = a_{m-1} * \overline{h}(2n), \quad (3)$$

$$d_{m,n} = \sum_p d_{m-1,p}\, g(p-2n) = d_{m-1} * \overline{g}(2n), \quad (4)$$

where $h(n)$ and $g(n)$ form a pair of conjugate mirror filters used at the analysis stage with $g(n) = (-1)^{1-n}h(1-n)$, $h(-2n) = \overline{h}(2n)$ and $g(-2n) = \overline{g}(2n)$ are synthesis filters, and n is the coefficient index. A delta time-scale feature $D(i)$ defined as the power difference between approximation and detail subbands is calculated for each speech frame as follows:

$$D(i) = \frac{1}{N_a}\sum_{n=1}^{N_a} E_{m,i}^2(n) - \frac{1}{N-N_a}\sum_{n=N_a+1}^{N} E_{m,i}^2(n). \quad (5)$$

where $N_a = \frac{N}{2^m}$ 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. The Teager energy operator (TEO) coefficients $E_{m,i}(n)$ are calculated from the wavelet coefficients $X_{m,i}(n)$ by the discrete form [11] as follows:

$$E_{m,i}(n) = X^2_{m,i}(n) - X_{m,i}(n+1)X_{m,i}(n-1). \quad (6)$$

The benefit of the TEO is a magnification of the amplitude difference between the wavelet coefficients in the approximation subband and the ones in the detail subbands. This improvement is certainly useful to extract a robust feature for the classification task in harsh environments, especially for the unvoiced frames. To be robust against strong and non-stationary noise, 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. Besides, the processed delta feature is further smoothed by a median filtering of five frames:

$$D_s(i) = \text{medfilt}\left(\frac{1 - e^{-2D(i)}}{1 + e^{-2D(i)}}\right). \quad (7)$$

3.2. Adaptive Threshold By Statistical Quantile Filtering

The enhanced feature values are compared with an estimated threshold related to the noise level to make a speech/non-speech decision. The statistical quantile filtering method which was proposed in [9] and has similar properties to the (SOSF) in [4] is used to estimate the noise threshold. Here, an optimized selection of quantile factors is proposed to have a better estimate of the noise threshold. We observe that, typically, the smoothed feature values $D_s(i)$ which represent a power relation between subbands stay at the noise level over a significant part of buffers ten seconds in length. Thus, the threshold adaptation is implemented as follows:

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

The quantile factor which had been selected experimentally as $q = 0.3$ from the range of $q = [0.0 \dots 1.0]$ provides a good estimate of the noise threshold [9]. However, we observe that the distributions between the number of speech frames and the number of non-speech frames varies for different buffers. Thus, the quantile factor $q(b)$ is apdated for each buffer to achieve a better noise threshold estimate. The method is based on a comparison of feature differences across five consecutive rank-sorted frames and a pre-determined level $\varepsilon = 10^{-3}$. The test is performed from the beginning of the buffer and is stopped when the difference is larger than this level. Then the quantile factor $q(b)$ is selected as:

$$q(b) = i', \qquad \text{if} \quad D_s(i') - D_s(i'-4) > \varepsilon \quad (8)$$

Figure 3 shows three different quantile factors $q(b)$ estimated more accurately than the constant quantile factor for three different buffers. Finally, the input frames 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. For smoothing of fluctuations resulting from strong non-stationary noise in the VAD outputs, the output sequence VAD(i) is smoothed by using the 100ms/200ms hangover scheme to bridge short voice activity regions, preserving only candidates with a minimal duration of 100ms, and being not more than

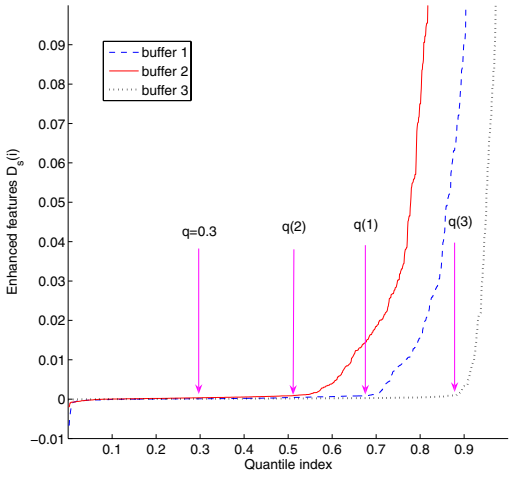

Figure 3: Adaptive quantile factors selected for different examples of frame buffers.

200ms apart from each other. This excludes talk-spurts shorter than 100ms and relabels pauses smaller than 200ms as speech. The impact of this rule is considered during our experiments.

4. Experiments and Discussions

Due to the lack of a noisy narrow-bandwidth database (e.g. air traffic voice database), the TIMIT database corrupted artificially by noise is used for experiments of building the VAD. The noisy fixed telephone SPEECHDAT-AT database and the clean WSJ0 database are used in experiments of designing the SV system. To simulate the conditions of ATC, all speech samples were band-pass filtered to a bandwidth of $300 - 2500$ Hz and down-sampled to a sampling frequency of 6 kHz. The proposed wavelet-based VAD algorithm (WVad) is compared with an energy-based VAD method (EVad) using the SQF technique developed in [8].

4.1. Evaluations on performance of VAD

From the TIMIT database, dialect speaking region 4, 80 continuous utterances (ca. 10100 frames of 32 ms frame length and 8 ms frame rate) were selected for experiments with the VAD. The dataset is divided into two subsets, 70% for training and 30% for testing. The speaker set contains 4 female speakers and 4 male speakers. The speech samples are artificially corrupted with additive white, car, and factory noise over the SNR range of [30, 20, 10]dB. Table 1 shows the average classification rates calculated from the confusion matrices over all frames of two classes speech/non-speech for the selected dataset, including noise-free and noisy recordings. The average classification rate (ACR) is calculated as a ratio between the sum of correctly accepted frames and the total number of tested frames over all classes.

For all three different types of noise, the WVad provides up to 2% lower error rate than the EVad method. Due to the highest complexity of factory noise which includes transient, colored, and non-stationary noise, the average performance over the considered SNR range derived by the wavelet-based method drops down by about 2.5% and 5% in comparison with the performance of the car and white noise cases, respectively. The

		White	Car	Factory
SNRs	Algs.	wo / w	wo / w	wo / w
30	WVad	93.68 / 95.47	92.77 / 94.67	91.29 / 92.95
	EVad	91.72 / 93.58	90.58 / 91.89	88.91 / 89.65
20	WVad	92.56 / 94.53	90.81 / 92.33	88.72 / 90.38
	EVad	90.87 / 92.11	88.45 / 90.53	86.03 / 87.93
10	WVad	90.23 / 92.85	87.13 / 89.71	84.82 / 86.32
	EVad	88.57 / 89.91	85.03 / 87.16	81.79 / 83.06

Table 1: The ACR (%) obtained for different algorithms (Algs.) without/with (wo/w) applying hang-over scheme and for different SNR and noise types.

application of the hang-over scheme really helps to increase up to 2.3% ACRs compared with the case of not using it.

4.2. Evaluations on the SV system

For this 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 the 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 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 $\lambda_{\text{UBM}}^{\text{Ge}}$. 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 was taken. 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.

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

Table 2: EER results derived from both databases for different VADs without (wo) and with (w) applying hangover scheme.

The equal error rate (EER) as special point in the Detection Error Tradeoff curve is used to measure performance. As reported in Table 1, for the noisy fixed line telephone recordings from SPEECHDAT-AT, the usage of both VAD methods improve 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 WVad than the EVad, the EER is reduced from 11.7% to 9 % without smoothing, and from 6.52% to 4.75% with smoothing. In addition, from the observed results, we discovered that the smoothing to bridge short pauses between speech frames helps to reduce the EER by 5.18% and 4.25% when using EVad and WVad, respectively.

5. Conclusions

In this paper, a robust voice activity detector is designed in the front-end unit for improving the performance of a narrow-bandwidth speaker verification system in harsh environments. The proposed wavelet-based VAD exhibits a very low-complexity as the detector uses only a single time-scale feature. Statistical quantile filtering is an excellent technique for estimating the noise level accurately, especially in case of non-stationary noise. The presented results show that the increases of the average classification rates indirectly reduce the equal error rates when using the wavelet-based VAD instead of the energy-based VAD. The hangover scheme with bridging rule which smooths the VAD output contributes to the EER reduction. For future work a realistic air traffic voice communication database will be tested with our SV system. Noise reduction may be embedded to remove fading channel noise in order to improve both, VAD and SV performance. Moreover, parameter fine tuning of the wavelet-based VAD method to minimize the EER will be investigated.

6. References

[1] Tanyer et al., "Voice activity detection in nonstationary noise," *IEEE TSAP*, vol. 8, pp. 478–482, 2000.

[2] Z. Xiong and T. Huang, "Boosting speech/non-speech classification using averaged mel-frequency cepstrum," in *Proc. of the Pacific-Rim Conf. on Multimedia*, 2002.

[3] B. F. Wu and K. C. Wang, "Voice activity detection based on auto-correlation function using wavelet transform and Teager energy operator," *Journal of Comp. Ling. and Chinese Lang. Proc.*, vol. 11, pp. 87–100, 2006.

[4] Ramirez, et al., "An effective subband OSF-based VAD with noise reduction for robust speech recognition," *IEEE TSAP*, vol. 13, no. 6, pp. 1119–1129, 2005.

[5] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE TSAP*, vol. 11, no. 5, pp. 466–475, 2003.

[6] ETSI, *ETSI ES 202 050 V1.1.3 Speech Processing, Transmission and Quality Aspects (STQ), Distributed speech recognition, Advanced front-end feature extraction algorithm, Compression algorithms*, 2003.

[7] I. Magrin-Chagnolleau, G. Gravier, and R. Blouet, "Overview of the 2000-2001 ELISA consortium research activities," June 2001, pp. 67–72.

[8] Neffe et al., *Speaker Classification*, chapter Speaker Segmentation for Air Traffic Control, Springer: LNAI, accepted for publication.

[9] Pham et al., "Low-complexity and efficient classification of voiced/unvoiced/silence for noisy environments," in *INTERSPEECH*, Pittsburgh, sept. 2006, pp. 661–664.

[10] Douglas A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," in *Speech Com.*, Sept 1995, vol. 17, pp. 91–108.

[11] J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *ICASSP*, 1990, vol. 1, pp. 381–384.