SPEAKER VERIFICATION FOR AIR TRAFFIC CONTROL

Michael Neffe 1, Franz Pernkopf 1, Tuan Van Pham 1, Horst Hering 2 and Gernot Kubin 1

1 Signal Processing and Speech Communication Laboratory
Graz University of Technology, Austria

2 Eurocontrol Experimental Centre, France
michael.neffe@tugraz.at, pernkopf@tugraz.at, v.t.pham@tugraz.at,
horst.hering@eurocontrol.int and g.kubin@ieee.org

ABSTRACT
In this paper a novel system of speaker segmentation has been
designed for improving safety on voice communication in air
traffic control. In addition to the usage of the aircraft ident-
fication tag to assign speaker turns on the shared commun-
ication channel to aircrafts, speaker verification is investi-
gated as an add-on attribute to improve security level effec-
tively for the air traffic control. The verification task is done
by training background and speaker models using Gaussian
mixture models. The front-end processing unit is optimized to
deal with small bandwidth restrictions and very short speaker
turns. To enhance the robustness of the verification system,
a cross verification unit is further applied. The designed sys-
tem is tested with SPEECHDAT-AT and WSJ0 database to
demonstrate its superior performance.

Index Terms— Air Traffic Control, Speaker Segmen-
tation and Verification, Gaussian Mixture Models

1. INTRODUCTION
There is a steady demand for increasing the security level in
Air Traffic Control (ATC) voice communication between con-
troller and pilots. In 2003 the Eurocontrol Experimental Cen-
tre (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 is proposed in [3]
by using a speaker verification (SV) system based on the AIT
information.

From Speaker Segmentation to Speaker Verification
The idea is that some behavioral biometric characteristics can
be extracted from the pilots’ voices and are automatically en-
rolled when an aircraft registers the first time to a control sec-
tor. At any later occurrence of the same AIT signature the
received speaker voice can be compared with the existing en-
rolled speaker dependent model to verify whether the speaker
had changed as proposed in [3]. Recapitulating, the AIT re-
duces the problem of distinguishing different speakers on the
party-line channel known as the speaker segmentation prob-
lem to a binary decision problem of claimant vs. imposter
speaker. This proposed concept secures the up and down link
of the ATC voice communication. An illustration of the pro-
solved solutions is shown in Fig.1. The first level shows the
talk spurts only. The AIT as shown in the second level in Fig.1
identifies beside the start of each sent voice message also the
transmitting source microphone. Based on this framework,
the level of security can be improved by embedding a SV
system which is depicted as the third level in Fig.1. At the
second and third level, the first numerical index determines
the aircraft number and the second the speaker. As an ex-
ample AC22 is not equal to AC21 which is the first speaker
assigned to aircraft AC2. Hence AC22 has to be verified as an
intruder. To satisfy the above demands the text-independent

Fig. 1. Coaction of AIT and SV on the party-line in air-
ground voice communication. TS is the abbreviation for an
arbitrary talk spurt, GC for a talk spurt originating from the
ground (i.e., ground control) and AC for a talk spurt originat-
ing from a certain aircraft.
by the VAD 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. Additionally for utterances with low confidence a cross verification unit can be used to increase recognition performance.

This paper is organized as follows: Section 2 investigates the restrictions arising from the transceiver equipment and the channel itself. In sec. 3 the system design is presented with all its processing units. A detailed description of the experimental setup and the databases is given in sec. 4 where also restrictions discussed in sec. 2 are considered. Experimental results and comparisons are discussed in sec. 5. The contribution finally ends with some conclusions in section 6.

2. VHF TRANSCEIVER EQUIPMENT AND ITS LIMITATIONS

After introducing the ATC security problem the signal conditioning and its effects for speech quality will be analyzed. Speech quality in ATC is mainly impaired by additive background noise (wind, engines) and by the radio transmission system and channel which limits the signal in bandwidth and causes distortions. The latter one is the main challenge for the SV system. ATC uses the amplitude modulation technique to transmit the signal over a Very High Frequency (VHF) channel with a channel spacing of 8.3 kHz. This yields in an effective bandwidth of only 2200 Hz in the range of 300 – 2500 Hz [4] for speech transmission. A thorough description of dominating effects like multipath propagation and a good deal more, degrading the transmitted signal can be found in [5, 6].

3. SYSTEM DESIGN

Based on the front-end processing unit speaker classification is performed. The basic design of this SV system consists of four phases as shown in Figure 2. In phase 1, gender dependent UBMs are trained. This step is carried out because of the lack of data to train speaker dependent models necessary to get independent to phonetic influences contained in the speech. 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 done.

3.1. Front-end Processing

The input speech signal is first fed to the VAD unit to divide speech from non-speech one. Based on the VAD output, features are extracted and then normalized to reduce influences not arising from the speaker shown in the dashed surrounded box in Fig. 2. These processed features are finally used for speaker classification.

![Fig. 2. Illustration of designing a SV system in 4 phases.](image-url)
rule proposed in [3]. To distinguish between speech signals and consistently high-level noise which results from the transmission channel itself during non-active communication periods, both methods are using to same long-term high-level noise detection [3]. For this the 1st discrete derivative of the correspondingly extracted feature sequence is calculated and used for detection.

3.1.2. Feature Extraction and Normalization

For each speech segment detected by the VAD method features are extracted separately as shown in Fig. 2. 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 filterbank 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 [9]. Altogether 42 features per frame are used.

In order to reduce the impact of channel dependent distortions, histogram equalization (HEQ) [10], a feature normalization method has been carried out as shown in Fig. 2. 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 that small to get sufficient statistical reliability of the data in each bin.

3.2. Classification

Here we use the GMM-UBM approach first introduced by [11]. In contrast to other GMM-UBM SV systems [9] we decided to train gender dependent UBMs which are finally not merged to one global UBM. Because of the lack of training data and the higher computational complexity, diagonal covariance matrices instead of full ones have been taken. 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 the retraining of the model to yield the final gender dependent UBM, we used 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 \( \lambda_{Spk} \) with its anti-hypothesis the gender UBM \( \lambda_{UBM_{Ge}} \):

\[
S(X) = \log L(X|\lambda_{Spk}) - \log L(X|\lambda_{UBM_{Ge}}) 
\] (2)

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 feature vectors of the first and second utterance to be investigated and \( \lambda_1 \) and \( \lambda_2 \) their adapted speaker models, respectively. If the following equation is satisfied

\[
S_{\lambda_2}(X_1) \& S_{\lambda_1}(X_2) > t 
\] (3)

i.e., both scores are above a threshold \( t \) and are verified to be from the same gender as defined in Eqn. 1, than it is assumed that both utterances are from the same person and thus are concatenated and used for verification. Figure 3 shows the region of insufficient confidence in the score distribution histogram. Intruders and true speakers are illustrated separately. The region of low confidence for our experiment as shown in this figure in the white box with dashed borders has been set to \(-1.8 \pm 0.2\) using the energy-based VAD.

![Fig. 3. Histogram and fitted Gauss curves for the score distributions of imposters (left) and true speakers (right). The rectangle with dashed borders illustrates the score region of low confidence.](image-url)
4. EXPERIMENTS AND DATABASE

The fixed telephone SPEECHDAT-AT database [12] and the WSJ0 database [13] 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 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 [3] 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 were 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 11.2.1 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 as well as for the reference speakers, 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.

5. RESULTS

To measure SV performance we use the Detection Error Trade-off (DET) curve and as special point in this curve the equal error rate (EER). DET-curve and EER are calculated using the score defined in sec. 3.2. As reported in Table 1, for the SPEECHDAT-AT database which consists of noisy 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 6.52% to 4.75% as illustrated in Fig. 4. Thus, by using the proposed WT-based VAD, we gain 23% relative improvement compared to the energy-based VAD. 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. More detailed information comparing both VADs and examining the influence of the bridging rule on EER can be found in [7].

![Fig. 4. Det-curve, as plot of false acceptance (FA) versus false rejection (FR) rate and EER point for the SV system without VAD (NoVad), energy-based (EVad) and WT-based VADs (WaVad).](image)

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. 5 we can report a reduction of the EER from 6.52 % to 6.12 %.

<table>
<thead>
<tr>
<th>EER [%]</th>
<th>NoVad</th>
<th>EVad</th>
<th>WaVad</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEECHDAT-AT</td>
<td>25.12</td>
<td>6.52</td>
<td>4.75</td>
</tr>
<tr>
<td>WSJ0</td>
<td>10.15</td>
<td>10.37</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 1. EER results derived from both databases for different VADs.

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 slightly 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.
6. CONCLUSION

To enhance security in the air traffic voice communication in this paper a speaker verification system is introduced as add-on to the aircraft identification tag. Our system was specially designed for the needs of air traffic control with respect to bandwidth restriction and talk spurt length. For speaker modeling the Gaussian mixture universal background model approach has been chosen with a front-end processing unit optimized for the air traffic control demands. Beside the feature normalization, the voice activity detection improves speaker classification performance under noise conditions as shown on the SPEECHDAT-A T database. Performance evaluations have been performed without VAD, with the energy-based and the WT-based VAD. The additional usage of the cross verification method has the ability to further reduce the EER. To reduce the impact of environmental mismatch application of noise reduction as pre-processing step and an adaptive threshold estimation algorithm for the cross-verification unit will be investigated.

7. ACKNOWLEDGEMENT

The authors kindly acknowledge the support of the EEC.

8. REFERENCES


