AMSE JOURNALS-AMSE IIETA publication-2017-Series: Advances B; Vol. 60; N°2; pp 283-303
Submitted May 25, 2017; Revised Sept. 21, 2017; Accepted Sept. 25, 2017

Speaker Identification Using Auditory Modelling and Vector Quantization

Konstantina Iliadi, Stefan Bleeck*

Institute of Sound and Vibration Research, University of Southampton, SO17 1BJ, UK
(bleeck@soton.ac.uk)

Abstract

This paper presents an experimental evaluation of different features for use in speaker identification (SID). The features are tested using speech data provided by the EUROM1 database, in a text-independent closed-set speaker identification task. The main objective of the paper is to present a novel parameterization of speech that is based on an auditory model called Auditory Image Model (AIM). This model provides features of the speech signal and their utility is assessed in the context of speaker identification. In order to explore the features that are more informative for predicting a speaker's identity, the auditory image is used within the framework of cutting it into rectangles. Then, a novel strategy is incorporated for the enrolment of speakers, which is used for specifying the regions of the image that contain features that make a speaker discriminative. Afterwards, the new speaker-specific feature representation is assessed in noisy conditions that simulate a real-world environment. Their performance is compared with the results obtained adopting MFCC features in the context of a Vector Quantization (VQ) classification system. The results for the identification accuracy suggest that the new parameterization provides better results compared to conventional MFCCs especially for low SNRs.

Key words

Auditory image model, Speaker identification, Feature extraction.

1. Introduction

Humans are considered to be fairly good at identifying speakers based on their voices. Automatic recognition systems are expected to do as well as humans but there is still a lack of robust speech characteristics that index an utterance as originating from one speaker rather than another.

The main goal in speaker recognition is to find measurable quantities that minimize within-speaker variability and simultaneously maximize between-speaker variability [1].

Generally, speaker recognition encompasses two fundamental tasks: speaker identification and speaker verification [2]. Speaker identification is the assignment of an unknown voice to one of the speakers known by the system and it is assumed that the voice must come from a fixed set of speakers. This task is often referred as closed-set identification. On the other hand, speaker verification refers to open-set identification because generally, it is assumed that the unknown voice may come from an impostor.

Regardless of the task, speaker recognition is a pattern classification task and consists of two procedures: training and testing. Training involves registration of speakers with the system. With registered speakers, speech data is matched with known patterns (speaker templates) and this process is called testing.

Both processes involve extraction of features from speech data, which makes the features critical to the classification process. To date, most speaker recognition systems use the MFCC (Mel Frequency Cepstral Coefficients) that are also used in speech recognition.

In this paper, we suggest an alternative to them called Auditory Image Model (AIM). The AIM is a visual representation of all the stages that a sound goes through once it enters the human auditory system. In Section II, the model, the feature extraction process and the chosen classification algorithm are described. The first experimental set, which is about identification in quiet conditions as well as specifying how the auditory model can be used in the best possible way, is presented in section III. Section IV consists of the evaluation of the AIM and MFCC in the context of speaker recognition in noisy conditions. The conclusions are presented in Section V.

2. Methodology

2.1 Auditory Image Model (AIM)

The auditory image model is a time-domain model of the signal processing stages in the hearing system associated with the ascending auditory pathway. Patterson et al. [1] first described

the auditory image as the simulation of the neural representation underlying humans' first conscious awareness of a sound.

The principle functions of AIM are to describe and simulate: 1) the basilar membrane motion (BMM) in the cochlea, 2) the neural activity pattern (NAP) observed in the auditory nerve and cochlear nucleus, 3) the identification of the peak times of the neural activity called strobe points, which are used to construct the auditory image and 4) the stabilized auditory image (SAI) that forms the basis of auditory perception.

Perceptual research on pitch and timbre indicates that at least some of the time-interval information in the NAP is preserved in the auditory image [4]. For that reason, Patterson *et al.* [1] supported that a continuous temporal average process cannot generally, simulate the auditory temporal integration, since averaging over time destroys the temporal fine structure within the averaging window. Furthermore, Patterson *et al*. [5] suggested that the fine structure of periodic sounds is preserved compared to the fine structure of noises.

As a result, they showed that this information could be preserved by, firstly, identifying peaks in the neural activity as it flows from the cochlea and measuring time intervals from these strobe points to smaller peaks. The final step is to form a histogram of these time intervals for each channel of the filter bank. This temporal integration process is referred to as strobed temporal integration (STI) and it stabilizes and aligns the repeating neural patterns of periodic sounds like vowels and musical notes [1,5]. The complete array of time interval histograms is the simulation of the AIM of an auditory image of the sound. Figure 1 shows the SAI for the vowel /ae/.

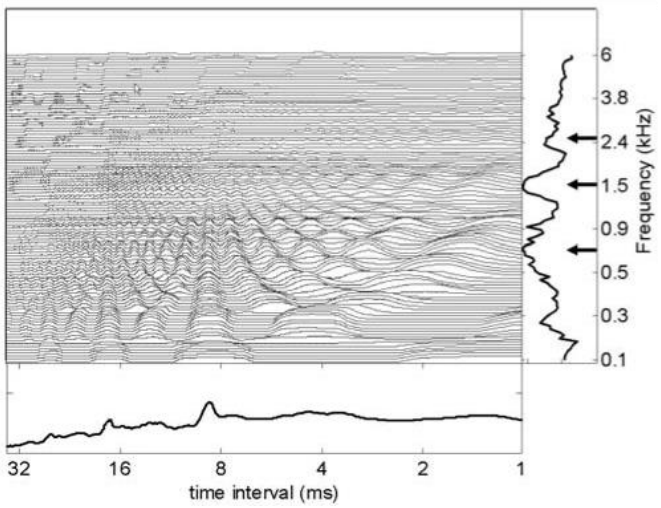

Fig.1. Main Panel: Stabilized Auditory Image of the Vowel /ae/.

The bottom panel is the temporal profile, which is estimated as the average over all channels for every point in time. The right panel is the spectral profile, which is the average over time for every channel. The arrows show the locations of formants. The peak at 9ms indicates the repetition rate of the sound. [4]

2.2 Feature Extraction

The auditory image can be used in the feature extraction stage in order to obtain feature vectors that represent it. These vectors can be processed for identifying patterns that typically appear in it.

Generally, the patterns can be identified at different positions in the image. The specific location of a pattern depends on the characteristics of the sound source. For pattern recognition, the information can be identified in smaller and larger scales in the SAI. At large scales, the temporal structure of the sound can provide information about the pitch whilst at smaller scales, there is information about the resonances following each pulse. The latter can be an indication for the vocal tract length (VTL) of a speaker.

Therefore, it is preferred to look for patterns in various locations and different scales of the SAI rather than the whole image. The process that is followed in order to recognize patterns is to define a set of overlapping rectangles of different scales that cover the whole frame.

At first, the initial rectangle size has been chosen to be 16 samples in the time interval dimension by 32 filter bank channels.

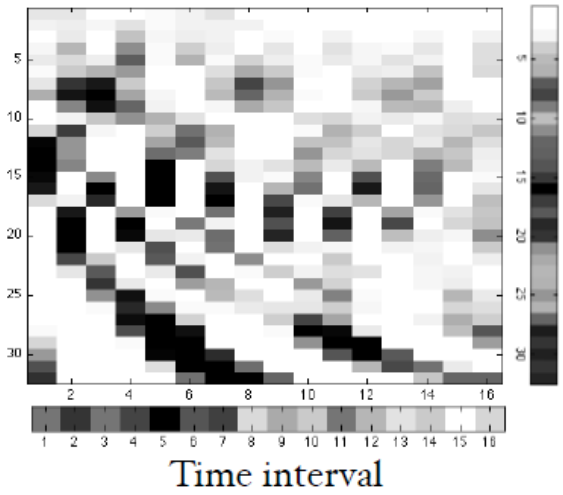

Fig.2. SAI Frame after the Down Sampling to 16 x 32 Pixels [6].

Then, from this baseline pair of box size, both dimensions were increased through multiplying them by powers of 2. The multiplication stops at the point where the dimensions of the largest box do not exceed the limits of the frame. For every pair of dimensions, the SAI space is tiled with boxes, starting at zero point in the time interval dimension. In the cochlear channel dimension, the box tiling occurs with a shift of half box width each time [6].

Afterwards, the content of each rectangle is independently processed with a specific concept named down sampling. The content of each box is reduced to the size of the smallest one (i.e. 16 samples in the time interval dimension by 32 filter bank channels). After this rescaling, the larger boxes are viewed at a coarser resolution. For further reduction of dimensionality, the margins of each rectangle are computed by averaging the elements over each of the two dimensions. Figure 2 shows a SAI frame that is downsampled to 16 x 32 pixels.

2.3 Modelling Framework

Vector Quantization (VQ) is a process of mapping vectors of a large vector space to a finite number of regions in that space. Each region is called a *cluster* and is represented by its centre (named *centroid*). A collection of all the centroids makes up a codebook.

Even though the codebook is smaller than the original sample, it still accurately represents a person's voice characteristics. In this procedure, a codebook is created for all the boxes that are extracted for the total number of SAI frames of every speech signal. The novel strategy that has been incorporated is analytically explained in the following section.

3. Speaker Identification in Quiet Conditions

In this section, we consider two different cases for the speaker recognition system in quiet environments. The first case is the identification for two databases of 30 and 180 speakers. The focus is on text-independent and closed-set identification, which assumes that the test case belongs to one among the registered speakers. The second one focuses on specifying the most discriminative features among speakers and obtaining a lower-dimensional feature representation from the SAI. In the next sections, the procedures and results for both cases will be presented in detail.

3.1 SAI-based System

In this section, the modules of the proposed system that uses AIM as a front-end are described. The architecture of the SAI-based system is presented in figure 3. The gammatone

filter bank that was used as the cochlea model for the auditory processing consisted of 64 frequency channels. Given the filter bank size, the SAI is cut into 154 rectangles and each box is reduced to a 48 – element feature vector with the down sampling process.

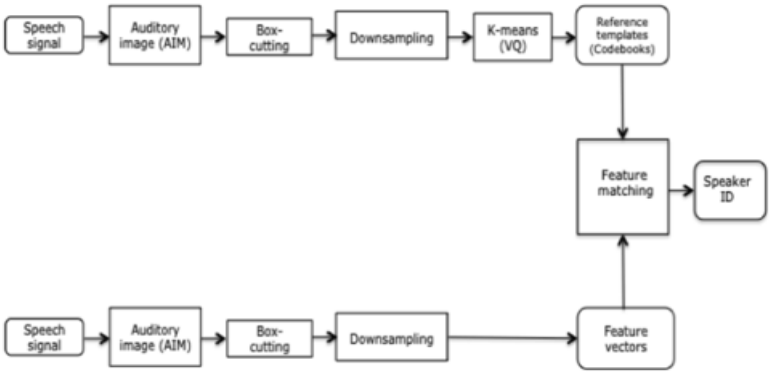

Fig.3. Architecture of the AIM-based SID System

After the transformation of the SAI frames into feature vectors, the speaker modelling step follows. This is achieved through the VQ, which is implemented using the K-means clustering algorithm. The number of centroids (or means) is chosen to be 64.

For the case of the 30-speaker corpus, the average duration of the training speech is 14.2ms, which results in 1420 frames. The number of boxes over the total number of frames is 1420 x 154 = 218680, i.e. 218680 feature vectors. For the corpus of 180 speakers, the average training speech duration is 20.7ms and the average total number of frames is 2070. Before the VQ, the feature dimensions are 318780 (2070 x 154 boxes).

After the VQ process (with K = 64), the speakers of both corpora are represented with a significantly reduced dimensionality of 64 x 154 = 9856 feature vectors. The final outcome of the enrolment session is a number of speaker models equal to the number of trained speakers. Each model consists of a number of codebooks equal to the total number of boxes (i.e. 154) for both cases that have been mentioned above. Each codebook has a size equal to 64 x 48 elements. The process is repeated as many times as the number of trained speakers.

Then, the testing phase occurs where the feature extraction stage is the same as in the training session. The features are extracted from every box of all frames as 48-element feature vectors.

For speaker matching, the concept is to see how well each codebook encodes the features of the target speaker through estimating the values that may reconstruct each frame of the test speaker using each one of the trained speaker models. To achieve that, the Euclidean distance is

computed, for every frame and every box, between every centroid in the codebook for that specific box and the current (test) feature vector for that box. For each frame, the minimum of these distances is the reconstruction value for that frame using the codebook for that specific box. Afterwards, the above process is repeated over the total number of frames and these values, for every box, can be averaged over all of the frames (i.e. the complete speech utterance). The latter results in the mean reconstruction value for every box.

Finally, the procedure is repeated for all of the trained speakers in the database. The speaker that is most likely to be the target speaker is the one that has the biggest number of boxes corresponding to the smallest average reconstruction value.

3.2 Baseline System

In this research study, the system that uses MFCC as a front-end is used as a comparison to the SAI-based system. The Mel-cepstrum is probably the most commonly used feature in speech recognition and has become the state-of-the-art method for speaker recognition as well.

For this system, the speaker modelling module consists of the same classifier as the proposed system that uses the K- means clustering algorithm. The difference between the two systems is that there is not a box-cutting module involved in the design of the MFCC-based system. Figure 4 presents its architecture

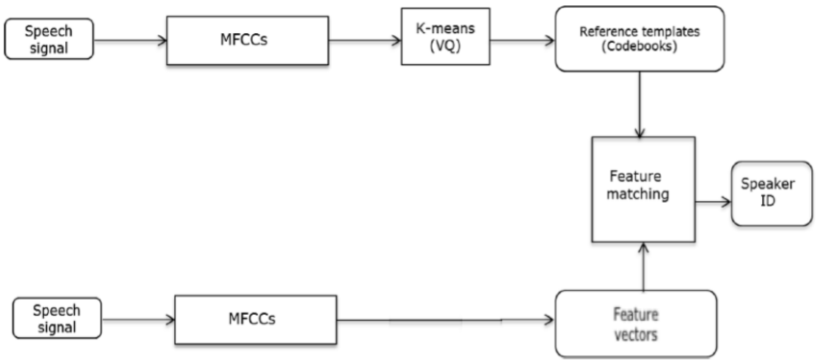

Fig.4. Architecture of the MFCC-based SID System

During the enrolment of speakers, 40 cepstral coefficients are extracted from every frame (of 25ms duration). The complete feature representation of the speech utterance consists of all of these vectors.

Afterwards, for speaker modelling, a codebook is learnt for every speaker over the total number of frames. The modelling framework is the same as in the SAI-based system so the centroids are equal to 64 and the codebook size for each speaker is 64 x 40 elements. When the

enrolment of speakers finishes, the end result is a number of reference templates equal to the number of enrolled speakers. Each template consists of only one codebook.

Subsequently, in the testing phase, the concept is to see how well each codebook encodes the features of the test speaker. So, for each frame, the 40-element test feature vector is computed and compared to each one of the 64 code words. This is based on computing the Euclidean distance between them and results in a matrix of distances to each centroid. In order to find the most representative centroid for reconstructing that frame using that codebook, the minimum distance between that centroid and the feature vector of the frame is computed. The same process is repeated for the total number of frames and the average of all of these distance values is estimated. The outcome is the mean value for reconstructing the whole speech utterance.

After repeating the process for all the speaker templates, the speaker matching is based on finding the template that corresponds to the smallest average value. This is the template that corresponds to the person that matches the unknown speaker.

3.3 Database Description

The speaker recognition experiments were performed on a multilingual corpus named EUROM1, which consists of recordings in 7 European languages [7]. For the purpose of this research work, a subset of EUROM1 was used and two different speech data sets were created.

The first corpus consists of 30 English speakers (12 females and 18 males), which are divided into 3 different groups of 10 people. The average length of the training and test speech is 14.2 and 15 seconds respectively.

The second corpus is 6 times larger than the first one and contains 3 groups of 60 talkers (30 females and 30 males in each group) in 3 languages (English, French and Swedish). The average training and test speech durations are 20.7 and 20.6 seconds respectively.

For all of them, the speech material is in the form of small passages that consist of 5 sentences. All speakers were prompted to read the same material and the recordings took place in an anechoic room. The speech signals have been pre-processed for pause removal.

3.4 Results

In this section, the identification performances of both the SAI-based and the baseline systems are presented in tables I and II. The SID accuracy was computed as the ratio of the number of correctly identified speakers to the total number of speakers that have been considered

for the testing phase. The error of the identification score is calculated as the standard error of the mean among the 3 different groups of speakers.

From the results, it is obvious that the proposed system achieves high accuracy levels. In the case of the small speech corpus, both systems achieve perfectly accurate identification. Also, when the number of speakers increases to 180, the performances of the two systems are similar. As expected, the SID accuracy decreases as the size of the speaker population increases.

After obtaining the identification results from the proposed system in quiet conditions, the process for finding the most informative SAI features is described in the next section and the results are presented.

Tab.1. Sid Accuracy (%) for the SAI-based and Baseline Systems (for the 30-speaker Corpus)

Corpus of 180 speakers		
System configuration	**SAI**	**MFCC**
SID Accuracy (%)	89.4	90.5
SID Error (%)	2	3.1

Tab.2. Sid Accuracy (%) for the SAI-based and Baseline Systems (for the 180-speaker Corpus)

Corpus of 30 speakers		
System configuration	**SAI**	**MFCC**
SID Accuracy (%)	100	100

3.5 Specification of Discriminative Features

In the previous section, the comparison between the proposed and the baseline systems shows that the SAI features can produce promising recognition results.

Yet, the issue of dimensionality is an important aspect in SID systems since it affects their computational efficiency. This experimental set is based on investigating the hypothesis that a subset of the extracted auditory features, which are more speaker-specific and have lower dimensionality, can be specified.

As previously described, each SAI frame has two dimensions, i.e. cochlear channel and time interval. The changes in the glottal pulse rate correspond to a change in the horizontal spacing of the vertical pitch ridges while a change in the resonance scale (formants) associates to changes in the vertical location of the resonance structure. Thus, the image separates, to a certain extent, the two types of information into its two dimensions.

In order to find the features that are more distinctive for every speaker, the VQ process that has been incorporated in the SAI-based system is used in a different context. As previously explained, the speaker matching happens when most of the boxes that are extracted from the SAI of the target speaker fulfil the criterion of minimization of the mean reconstruction value from the codebooks, which correspond to those boxes, of a specific speaker template. Inversely, when the speakers do not match with each other, these reconstruction values should be maximized. If these maximum values are estimated, the boxes that correspond to them can be specified. The features that are related to those boxes are the discriminative ones among all speakers. Additionally, the position of these boxes can indicate which areas of the image are more informative for speaker identification. This knowledge can also be useful for developing the box-cutting process from its initial version.

For this experiment, the data sets and the speech material are the same that were used before. The procedure was repeated 3 times for each group of the 10 and 60 speakers. Figures 5 and 6 show where these areas are located on the image for all speakers.

In these figures, the x-axis is the time interval dimension while the y-axis is the frequency channel dimension of the SAI. The plotted rectangles are those that have been specified by the combination of the box-cutting and VQ procedures and fulfil the maximization criterion. Every box has a size of 32 frequency channels and 16 samples in the time interval dimension.

From the figures, it is apparent that the most informative regions are located at approximately the first 10ms of the SAI in terms of the time interval dimension. Commonly, the first glottal pulse that forms the first pitch ridge lies in that time span. The glottal pulses are known to be produced by the vocal cords in the larynx and excite resonances in the vocal tract beyond the larynx. The larynx varies among people of different gender and age and so does their pitch that is considered to be a source of individuality in the voicing mechanism.

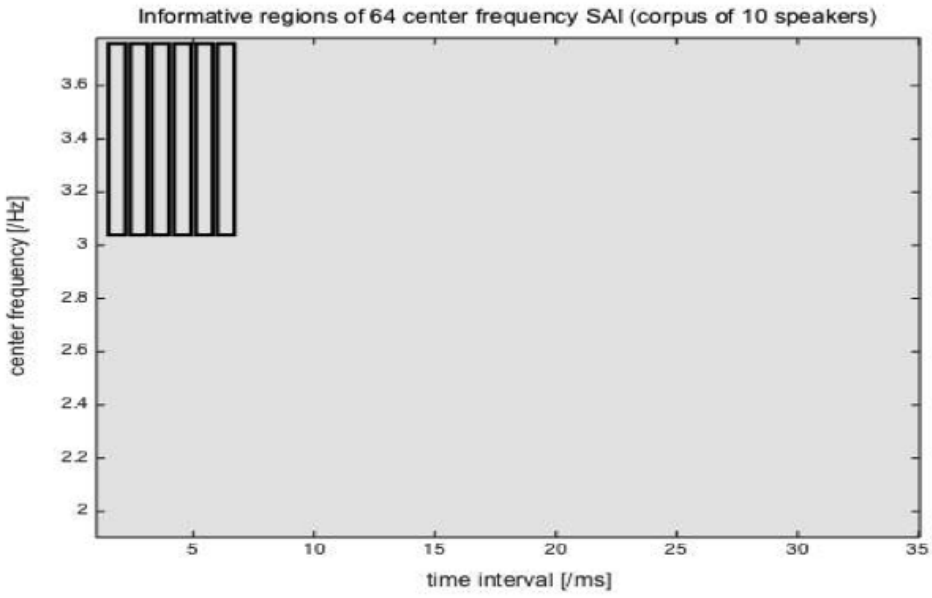

Fig.5. Specification of the Informative SAI Areas (for the 3 Trials Using 10 Speakers Each Time).

The 6 Boxes Cover the Part of the Filter Bank above 1KHz and the Area between 1.6 and 6.4ms.

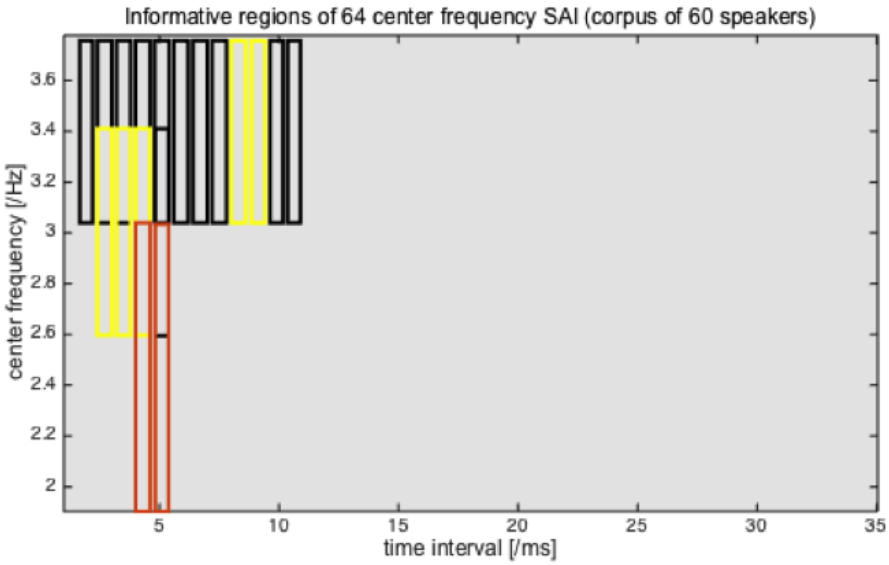

Fig.6. Specification of the Informative SAI Areas (for the 3 Trials Using 60 speakers Each Time). The 18 Boxes Cover Different Part of the Filter Bank (Low and High Frequencies) and Extend up to 11.2ms. The Black Boxes are Specified in All Trials While the Yellow and Orange Ones Are Discriminative as Well for the First and Second Trial Respectively.

Furthermore, another element that varies independently from the glottal pulse rate (or pitch) is the resonance scale (formants). Even though a person can alter his/her pitch, the resonances of one's vocal tract will not change because the anatomy remains the same. As a result, it seems that

the formants are a more critical feature for identifying a person from one's voice. Also, the boxes that are located beside each other horizontally contain a segment of the structures that are a result of resonances. Thus, this type of information is speaker-dependent.

Additionally, a very interesting finding of the VQ process is that the high frequency range of speech may contain meaningful speaker information. More specifically, the location of the boxes in figures 5 and 6 above 1KHz indicate that the high frequencies should not be overlooked. Consequently, higher formants can be distinctive and should be combined with the lower ones in order to make a decision for correct identification.

In conclusion, the incorporation of this new training strategy and the alternative use of it makes it possible to extract the notable information that makes a speaker more discriminative compared to others. Overall, it seems that the area of interest converges at, approximately, the first 10ms of the auditory image.

Finally, the above results support the hypothesis that perceptual differences about speakers can rely on lower-dimensional features. The latter will be important for the further development of the SID system design in the next set of experiments.

4. Noise-Robust Speaker Identification

The focus in this part of the study, is firstly on improving the existing system design through using the obtained knowledge from the previous experiments. Then, the objective is to assess the auditory features under the presence of interfering sounds, which is a common challenge for SID systems. The improvement of the existing box-cutting module as well as the results from the comparison between the SAI and the MFCC parameterization are presented as follows.

4.1 SAI-based System

First of all, the system architecture is the same as it has been presented in figure 3 as well as the operations and parameters involved in the auditory model.

Furthermore, the first objective of this experimental part is to modify the box-cutting process in order to create a more computationally efficient system. This is achieved through taking into account the results of the specification of the most distinctive features. As mentioned previously, the position of the rectangles that contain these features converge in, approximately, the first 10ms of the time interval dimension. Also, it is important to consider that most of the boxes are placed alongside each other and if they are added up together, they cover the image on a larger scale. Therefore, this leads us into choosing to segregate this specific region of the image where

all the boxes are located. The latter helps in eliminating the substantial redundancies that exist in the SAI.

In consequence, the image is cut into one rectangle that covers the whole filter bank and reaches up to 12.8 ms (256 time samples) in the time interval axis. The choice of this size is made on the basis of the dimensions of the rectangles in the box-cutting process being powers of 2. So, for computational convenience, we choose the area that extends up to 256 time samples (which is a power of 2) instead of 200 samples (i.e. 10 ms). In terms of the frequency dimension, the selection of all the frequency bands is based on trying to preserve the spectral formation without disassembling structures that are caused by resonances.

After this modification, the end result is a box of 64 frequency channels and 256 samples for every frame and it is shown in figure 7. After cutting these boxes for the total number of frames, all of them are downsampled as before into the smallest box of 32 x 16 pixels and reduced into 48 values each. The final outcome of the developed feature extractor is a number of feature vectors, which consist of 48 elements, equal to the number of SAI frames.

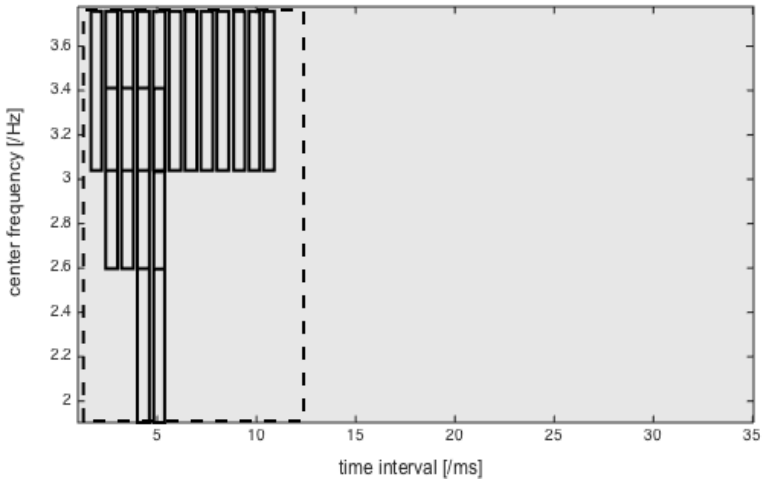

Fig.7. Selection of the SAI Area that Contains the Speaker-Specific Features. this Box (Dashed Line) Covers the 64 Channels of the Filter Bank and Extends up to 12.8ms in The Time Interval Axis.

The step that follows the feature extraction is the speaker modelling, which creates the trained speakers' reference templates, and it is implemented through the K-means clustering algorithm. For every speaker, one codebook is created for that one box that is cut over the complete number of frames. More specifically, all the 48-element feature vectors are concatenated in order to have the representation of the entire speech signal. After the VQ, each speaker is represented by a codebook of 64 code words since the number of centroids remains the

same for this experiment as previously. As a result, the outcome of the enrolment session is a number of speaker templates equal to the number of trained speakers that consist of a codebook with 64 x 48 features. The latter is a similarity with the baseline system since the MFCC-based system uses the single codebook approach with 64 x 40 elements.

Afterwards, the speaker testing occurs with the same feature extractor as shown in figure 3. At this stage, the same modification in the box-cutting takes place as it is presented in figure 7. For each box of every frame, the result of the down sampling is a vector with 48 features. After gathering all the feature vectors of the target speaker's speech, the concept is again to see which reference template encodes them better. To achieve that, the same steps, which were followed in section III, for estimating the Euclidean distances between the centroids and the feature vectors, are repeated. The criterion for speaker matching is the minimization of the average value that reconstructs the test speech signal using a specific codebook.

In conclusion, the system development that has been described is a salient contribution since the feature dimensionality is reduced while complex structures in the SAI data are being captured. Also, the use of a single codebook, compared to the multi-codebook approach of its initial design, makes a difference to the computational complexity of the system.

Lastly, the performance of the modified SID system will be evaluated through comparing it with the baseline system that has the architecture described in section III (figure 4).

4.2. Database Description

The second part of this research work comprises of two types of evaluation. At first, the hypothesis that the SAI features can be more noise-robust compared to the MFCC is tested. In this set of experiments, the data sets are part of the same multilingual EUROM1 corpus that has been used in the first part of this study. The system is again randomized in terms of the groups of speakers so that the variability of SID accuracy is estimated. The speech corpora of 30 (3 groups of 10) and 180 (3 groups of 60) talkers have been used here as well. The characteristics of the training and test speech material are the same as before.

Nevertheless, the main difference is that the training speech is clean whereas the test speech utterances are mixed with babble noise of 8 talkers at various SNR levels from -5 to 10 dB, at 5 dB intervals. For these identification experiments, each test case was matched against all speakers and the closest one was taken as the result.

Then, our next hypotheses are that the duration of the training and test speech material can affect the levels of accuracy. The former is tested through varying the duration of the training

speech while the test speech length remains constant. Inversely, in the latter, the training speech duration remains constant whilst the test speech varies. For these experiments, the database consists of 60 English speakers of EUROM1 that are divided into 6 groups (in order to test the variation of the accuracy). In both cases, clean speech is used for the enrolment session whilst the test speech is mixed with the same type of babble noise that was used before but only at 0 dB SNR.

4.3 Results

Firstly, the novel SAI parameterization is compared to standard MFCC. The speaker modelling framework of VQ with 64 centroids remains constant for both systems throughout all of the experiments in order to focus on the feature sets. The SID accuracy was computed in the same way as in the previous experiment. The error of the identification score is estimated as the standard error of the mean.

Figure 8 plots the SID accuracy (%) against the SNR levels for the corpus of 30 speakers. In general, the results indicate that the auditory features provide satisfying accuracy for all noise levels. Additionally, they perform significantly better for -5 and 0 dB SNR (t-tests with p-value equal to 0.0352 and 0.0301 respectively).

Specifically, the accuracy at -5 dB SNR is equal to 50% compared to 30% obtained by the MFCCs (with a 5.77% error for both cases). Also, for 0 dB SNR, this configuration achieved on average 90% identification (with a 10% error), while the baseline system achieved 60% recognition (with a 5.77% error). Interestingly, as the noise level decreases, performance reaches a point of saturation. For 5 dB SNR, the two systems have equal identification accuracy of 93.33% with an error of 3.33%. Maximum accuracy is attained for 10 dB SNR, where the scores are quite similar. MFCCs achieve 100% identification whilst the SAI reaches 96.66% with an error of 3.33%.

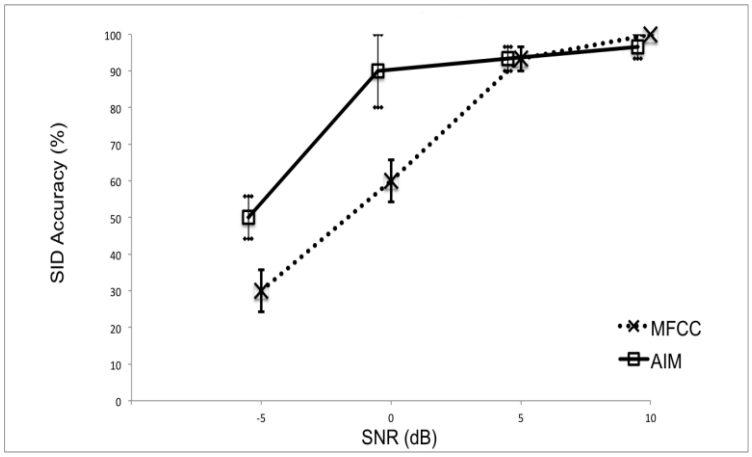

Fig.8. Speaker Identification (SID) Accuracy (%) of the SAI-based and MFCC-based Systems for the Corpus of 30 Speakers Using Multi-talker Babble Noise. the Error Bars Represent the Standard Error of the Mean (Among the Levels of SID Accuracy of the 3 Subsets of 10 Speakers)

Figure 9 presents the results for the identification accuracy against the 4 SNRs for the larger speaker population. From the figure, it is clear that the hypothesis for the noise robustness of the SAI features can be justified despite the change in the number of speakers.

In particular, for 0 dB SNR, it is noteworthy that the recommended configuration reaches on average 55% recognition (with a 7.26% error) compared to 35% (with a 7.8% error) obtained by the MFCCs. For even more noisy conditions, reflected by -5 dB SNR, the average identification is almost 5.5% higher for the auditory features (17.22% with 6.82% error).

As the SNR increases, there seems to be a saturation of performance with comparison to the baseline system. Yet, for 5 dB SNR, there is still better identification score (71.66% with 8.55% error) whilst for 10 dB SNR, there is convergence of the outcomes of both systems, i.e. 81.11% with 8.94% error for the SAI and 82.77% with 4.33% error for the MFCCs.

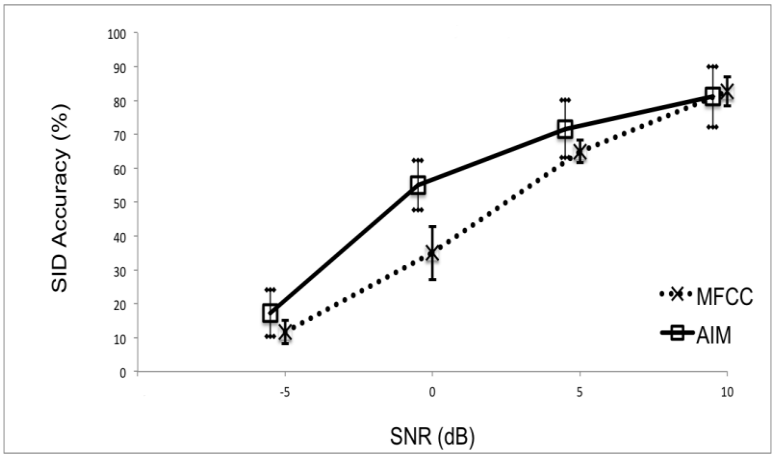

Fig.9. Speaker Identification (SID) Accuracy (%) of the SAI-based and MFCC-based Systems for the Corpus of 180 Speakers Using Multi-talker Babble Noise. The Error Bars Represent the Standard Error of the Mean (Among the Levels of SID Accuracy of the 3 Subsets of 60 Speakers)

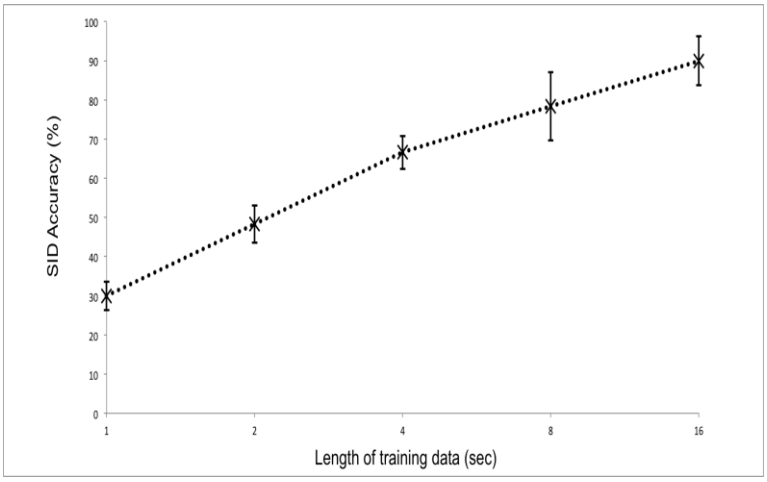

Fig.10. Speaker identification (SID) accuracy (%) of the SAI-based system for varying training speech duration. The error bars represent the standard error of the mean (estimated as the error among the levels of SID accuracy of the 6 subsets of 10 speakers)

In the second part of this evaluation, the hypothesis that larger amounts of training data result in better recognition rates is investigated. As previously mentioned, the test speech length remains the same while the training speech is doubled for every trial and varies from 1 to 16 secs (which is the maximum average duration of this corpus). Speakers are modelled using the 64-means clustering algorithm. Figure 10 shows the results for the SID accuracy of the SAI – based system against the length of the training speech material.

As expected, the identification accuracy gets better as the length of the training speech increases. This generally happens since more training data result in obtaining more reliable estimates of the speaker models. Even though the codebook size (64 code words) is not very large, it is remarkable that the accuracy level for 1 sec of training data is 30% (with 3.65% error), 48.3% (with 3.65% error) for 2 secs and 66.6% (with 4.22% error) for 4 sec.

Additionally, it appears that for these cases, there is a constant relationship between the two variables since the accuracy improves up to 18.3% for every doubling of the speech duration. This observation is also valid for the speech durations of 8 and 16 secs, where the recognition rate is 78.3% (with 8.7% error) and 90% (with 6.3% error) correspondingly. For both cases, the

performance improves up to 11.7%, which is less than the previous upgrade, but still remains the same.

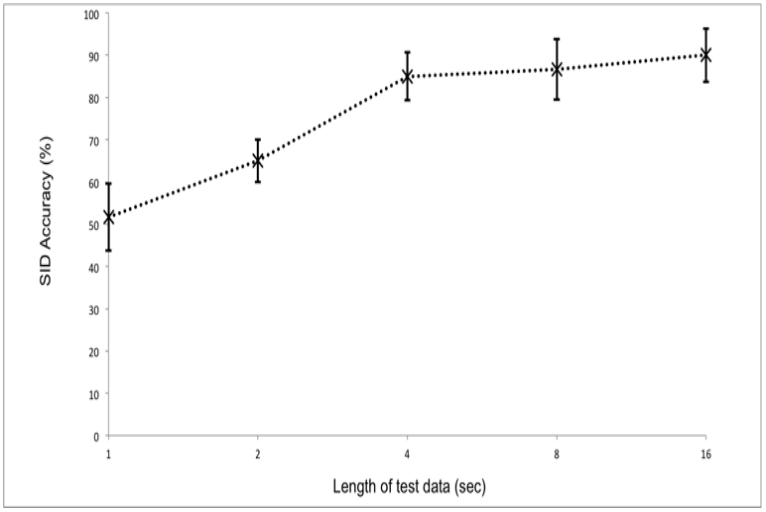

Fig.11. Speaker identification (SID) accuracy (%) of the SAI-based system for varying test speech duration. The error bars represent the standard error of the mean (estimated as the error among the levels of SID accuracy of the 6 subsets of 10 speakers)

Finally, in the third part of this evaluation, the hypothesis about the effect of the length of test speech on the identification score is tested. Usually, longer test data are expected to produce better accuracy levels since more information can be extracted to represent the target speakers. In this case, the speech used during enrolment does not change and the test speech segments range from 1 to 16 sec. For each trial, their duration is doubled. The speaker modelling framework is the same. The results for the SID accuracy against the length of the test speech are summarized in figure 11.

From the results in the figure above, it seems that the performance becomes better as the test speech length increases. When the test utterances last for 1 and 2 secs, the average identification accuracy is 51.6% (with 7.9% error) and 65% (with 5% error) respectively. Importantly, the biggest improvement of performance occurs for the speech length equal to 4 secs, which is 20% on average, and the SID accuracy reaches 85% (with 5.6% error).

Additionally, for the length of 8 secs, the accuracy is almost similar to that of 4 secs and equal to 86.6% (with 7% error) while it slightly improves up to 90% (with 6.3% error) for the maximum speech duration of 16 sec. Lastly, it is worth mentioning that after the duration of 4 secs, there seems to be a saturation of the recognition score.

Conclusion

In this paper, the applicability of an auditory model, named Auditory Image Model, on a text-independent speaker identification system has been discussed. Our research work consisted of two main sets of experiments.

In the first one, the identification task was conducted in quiet conditions for two speaker data sets of different sizes. The system performance was compared to a baseline system using the MFCC parameterization. The results suggest that the features that are extracted from the auditory image can produce high recognition cores similar to those obtained by the benchmark.

Furthermore, the second part of this experimental set investigated the hypothesis that it is possible to retrieve a subset of the auditory features that is more speaker-specific. The latter was achieved with the incorporation of a novel strategy during the enrolment of speakers that combines the method of box-cutting with VQ. The concept was to analyse the content of every box of the image independently through creating codebooks for each one of them. This way of yielding codebooks resulted in specifying the most informative regions of the SAI that indicate features that are more discriminative for speakers.

After the specification of these areas of the image, the boxes converge to the area up to, approximately, 10 ms in terms of the time interval dimension. With regard to the frequency dimension, the rectangles cover the whole filter bank or parts of it. Since the first pitch ridge lies in that SAI region, it appears that pitch is one source of individuality. Also, the boxes contain part of the structures that have been created as a result of the resonances of the vocal tract (lower and higher formant frequencies), which is a very important characteristic of the anatomy of a person.

Lastly, the first glottal pulse is usually included in that time span and the shape of it can affect the speaker's voice quality. As a result, it appears that this SAI region can provide information about the characteristics of a speech signal that are speaker-dependent.

Overall, it seems that the benefit of the SAI approach is that the signal is converted into a two-dimensional representation that makes it possible to segregate the glottal pulse rate from the resonance structure of the vocal tract. This allowed us to specify the characteristics mentioned above that make a speaker more discriminative compared to others.

In the second experimental set of this study, the results of the previous part were used in order to deal with the issue of feature dimensionality. The existing configuration of the box-cutting module was modified so that it includes the informative auditory features. This procedure is important since there were substantial redundancies in the image and it was essential to try and find a denser representation of the signal with reduced data dimensionality. Additionally, after the

speaker modelling part, each speaker template consisted of a single codebook, which is different from the multi-codebook approach that was used in the initial version of the system. Consequently, the new representation makes a good comparison with the MFCC features since their dimensionality is similar and it is computationally efficient given that the features are extracted in much less time for both the training and testing phases.

Then, the robustness of this novel representation was investigated in noisy conditions that simulate a realistic environment for two different speaker databases. For both cases, the results suggest that the auditory feature vectors lead to much better performance, i.e. higher SID accuracy, compared to the MFCC-based system especially for low SNRs.

Overall, it seems that one characteristic of the SAI that is key to noise robustness is the representation type of the auditory image, which has the benefit of combining different types of information to a certain extent. The first kind of information is the use of the temporal fine structure at the output of the filter bank. This results in the SAI preserving the fine timing information whereas the MFCC retain the spectral envelope.

Another important element of the auditory model is that the image contains the relative magnitudes of all frequency bands. At the same time, it includes the specific positions of the frequency areas with high magnitudes that associate to resonances of the vocal tract. This trait of the SAI is one of the reasons behind its robustness for distorted speech, since more noise can be tolerated around the spectral peaks.

In the final part of these experiments, two hypotheses were tested about the durations of the training and test speech segments influencing the SID accuracy levels. As expected, the system performance improved as the length of the speech utterances increased in both cases. Nevertheless, it is notable that the proposed system achieved very satisfying recognition scores for relatively short training and test speech utterances.

In conclusion, it seems that the derived features are promising and merit the attention of the speaker identification and verification community for consideration in further work.

Acknowledgements

The authors would like to thank Tom Walters for the feedback on the box-cutting procedure used in the design of the proposed SID system.

References

1. R.D. Patterson, M.H. Allerhand, C. Giguere, Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform, 1995, JASA, vol. 98, pp. 1890–1894.
2. D.A. Reynolds, An overview of automatic speaker recognition technology, 2002, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP'02), vol. IV, pp. 4072-4075.
3. P. Rose, Forensic Speaker Recognition, 2002, Taylor and Francis, Inc., New York.
4. S. Bleeck, Ives, T. & Patterson, R.D. (2004). Aim-mat: The auditory image model in MATLAB. Acta Acustica, 90, pp. 781–787.
5. R.D. Patterson, K. Robinson, J. Holdsworth, Complex sounds and auditory images. In: Auditory physiology and perception. Y. Cazals, L. Demany, K. Horner (eds.), 1992, Pergamon, Oxford, pp. 429–446.
6. R.F. Lyon, M. Rehn, S. Bengio, T.C. Walters, G. Chechik, Sound retrieval and ranking using sparse auditory representations. Neural computation, 2010, vol. 22, pp. 2390–416.
7. D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, J. Zeiliger, EUROM- A spoken language resource for the EU, 1995, Eurospeech'95 Proceedings of the 4th European Conference on Speech Communication and Speech Technology, Madrid, Spain, 18-21 September. vol. 1, pp. 867-870.
8. X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, S. Shamma, Linear versus Mel-frequency cepstral coefficients for speaker recognition, 2011, Proceedings of IEEE Workshop on ASRU, pp. 559-564.