Voice-Based Gender Identification in multimedia applications

Hadi Harb, Liming Chen

Voice-Based Gender Identification in Multimedia Applications

HADI HARB, Hadi.harb@ec-lyon.fr
LIMING CHEN liming.chen@ec-lyon.fr
LIRIS CNRS FRE 2672, Dept. Mathématiques Informatique, Ecole Centrale de Lyon, 36 avenue Guy de Collongue, 69134 Ecully Cedex, France

Abstract. In the context of content-based multimedia indexing gender identification based on speech signal is an important task. In this paper a set of acoustic and pitch features along with different classifiers are compared for the problem of gender identification. We show that the fusion of features and classifiers performs better than any individual classifier. Based on such conclusions we built a system for gender identification in multimedia applications. The system uses a set of Neural Networks with acoustic and Pitch related features.

90% of classification accuracy is obtained for 1 second segments and with independence to the language and the channel of the speech. Practical considerations, such as the continuity of speech and the use of mixture of experts instead of one single expert are shown to improve the classification accuracy to 93%. When used on a subset of the Switchboard database, the classification accuracy attains 98.5% for 5 seconds segments.

Keywords: content-based audio indexing, Piecewise Gaussian Modeling, mixture of neural networks

1. Introduction

Automatically detecting the gender of a speaker has several potential applications. In the context of Automatic Speech Recognition, gender dependent models are more accurate than gender independent ones. Hence, gender recognition is needed prior to the application of one gender dependent model (Acero and Huang, 1996; Neti and Roukos, 1997). In the context of speaker recognition, perfect gender detection can improve the performance by limiting the search space to speakers from the same gender. In content based multimedia indexing, the speaker's gender is a cue used in the annotation. Also, Gender dependent speech coders are more accurate than gender independent ones (Marston, 1998; Potamitis et al., 2002). Therefore, automatic gender detection can be an important tool in multimedia signal analysis systems.

Several acoustic conditions exist in audio-visual data: compressed speech, telephone quality speech, noisy speech, speech over background music, studio quality speech, different languages, and so on. Clearly, in this context, a gender identification system must be able to process this variety of speech conditions with acceptable performance.

In this paper we propose a gender identification system based on a general audio classifier. The proposed technique doesn't assume any constraint on the speech quality or segment lengths, in contrary to the existing techniques.

2. Related work

Motivated by different applications, several works have focused on voice based gender detection in the literature. Konig and Morgan (1992) extracted 12 Linear Prediction coding Coefficients (LPC) and the energy features every 500 ms and used a Multi Layer Perceptron as a classifier for gender detection. They reported an 84% of frame based accuracy on the DARPA resource management database (Price et al., 1988). The database is a collection of clean speech recorded from 160 speakers in American English. Vergin and Farhat (1996) used the first two formants estimated from vowels to classify gender based on a 7 seconds sentences reporting 85% of classification accuracy on the Air Travel Information System (ATIS) corpus (Hemphill Charles et al., 1990) containing specifically recorded clean speech. Neti and Roukos (1997) used a simple pattern matching approach where the acoustic observation of a speech utterance is first decoded into phonemes and the Euclidian distance is calculated between the observation and the recognized male and female phoneme models. The model with the lowest distance is chosen to detect the gender of the speaker. The results on the ATIS corpus are 62% of accuracy for sentences from 3 to 8 seconds. However, when using a general GMM approach to model each gender's acoustic vectors, Neti et al. report in the same paper classification results of 95% precision rate on the same sentences of 3 to 8 seconds. In order to deal with the problem of gender normalization of speech (Jung et al., 2002) used pitch detection based on the simple Average Magnitude Difference Function (AMDF) in gender identification on the DARPA TIMIT speech corpus. Tzanetakis and Cook (2002) followed a general audio classifier approach using Mel Frequency Cepstral Coefficients (MFCC) features and Gaussian Mixture Models (GMM) as a classifier. When applied to gender identification in a multimedia indexing context, the results are 74% of classification accuracy with three classes, male, female and sports announcement. Slomka and Sridharan (1997) used a combination of a pitch-based approach and a general audio classifier approach using GMM. The reported results of 94% are based on 7 seconds files after silence removal on the OGI and the Switchboard telephone speech databases (Muthusama et al., 1992; Godfrey et al., 1992).

Hidden Markov Models were also used for gender identification. For each gender, one HMM speech recognition engine is trained. The gender dependent models are used to decode a test speech signal. The model with higher likelihood is chosen as a cue for the gender of the speaker (Huang et al., 1991). Parris and Carey (1996) combined pitch and HMM for gender identification reporting results of 97.3%. Their experiments have been carried out on sentences of 5 seconds from the OGI database. Some studies on the behavior of specific speech units, such as phonemes, for each gender were carried out (Martland et al., 1996).

This overview of the existing techniques for gender identification shows that the reported accuracies are generally based on sentences from 3 to 7 seconds obtained manually. Moreover, some papers use sentences from the same speaker. This assumption holds for speech data where speaker boundaries are known in advance, such as the case of telephone speech. Although the automatic segmentation of a continuous stream of speech, such as the one existing in multimedia documents, is feasible, it still constitutes an important problem for the multimedia indexing community (Viswanathan et al., 2000; Delacourt and Wellekens,

2000). Generally Equal Error Rates (EER) are of the order of 15% depending on the nature of the documents (broadcast news are "easier" to segment than meeting recordings for example).

In several studies, some preprocessing of speech is also done, such as silence removal or phoneme recognition. While silence can be efficiently detected for clean speech, it becomes problematic for noisy speech. Phoneme recognition on the other hand adds an additional complexity layer to gender identification systems by the need of training phoneme recognizers on a specific speech corpus.

3. Gender identification in multimedia applications

In the context of multimedia applications gender identification has specific characteristics that make it different from the gender identification in other applications.

(1) The speech signal in multimedia data is recorded from a variety of different sources. For instance, indoor, outdoor, and telephone speech are frequent in multimedia data. Therefore, a gender identifier for multimedia applications must be robust to channel and acoustic condition changes. In the systems proposed in Parris and Carey (1996) and Slomka and Sridharan (1997) it is supposed that the speech is recorded from the telephone network and hence restricting the use.
(2) Several audio coding and compressing techniques are used and hence implying the robustness to audio compression techniques.
(3) The speech is Multilanguage implying that a gender identifier in this context must be language independent. This assumption makes the use of HMM-based gender identification systems not effective since they are phoneme-related.
(4) The speech is a continuous stream with no *a priori* known boundaries at each speaker turn. That is, it is not guaranteed that relatively long term audio segments contain speech from the same speaker or gender. On the other hand the time precision in multimedia applications is crucial, it is considered that 2 seconds is a maximum delay in such applications.

The system described in this paper makes use of several features and classifiers and fulfills the above requirements for gender identification.

4. The general audio classifier approach

Gender identification is a classic pattern recognition problem with two classes. In this section we describe the main features used for gender identification and we propose a new set of perceptually motivated features arguing that they are a better alternative to the classically used acoustic features. Moreover, several classifiers are described and a comparison is carried out.

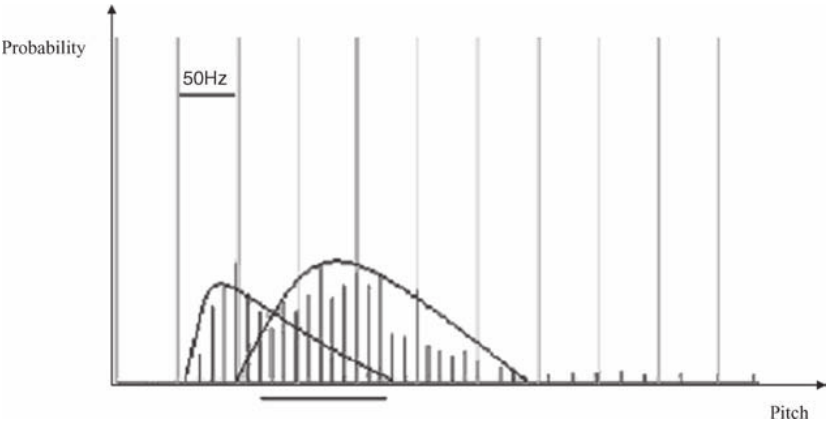

Figure 1. Pitch Histogram for 1000 seconds of males (lower values) and 1000 seconds of females' speech (higher values). We can see the overlap between two classes.

4.1. The features

4.1.1. Pitch features. The pitch feature is perceptually and biologically proved as a good discriminator between males' and females' voices. However, the estimation of the pitch from the signal is not an easy task. A good estimate of the pitch can only be obtained for voiced portions of a clean non-noisy signal (Hess, 1983; Shimamura and Kobayashi, 2001; Ross et al., 1974). Moreover, an overlap of the pitch values between male's and female's voices naturally exists, hence intrinsically limiting the capacity of the pitch feature in the case of gender identification, figure 1.

In our experiments we use a standard algorithm for pitch estimation based on the autocorrelation function in the time domain for windows lengths of 100 ms. The algorithm estimates the pitch for the voiced portions of the signal. Voiced portions are detected using the Zero Crossing Rate and the accepted pitch values are between 20 and 600 Hz.

In the presented system the gender identification is performed on frames of 1s, called Integration Time Windows (ITW). Therefore, in each ITW window 10 pitch values can be extracted, called All Pitch values (AP). Consequently, the mean (Mean Pitch, MP) and the minimum of the pitch (minP) values are obtained for each ITW.

4.1.2. Acoustic features. Short term acoustic features describe the spectral components of the audio signal and are generally extracted at 10 ms rate. Fast Fourier Transform can be used to extract the spectral components of the signal. Further filtering based on the perceptually motivated MEL scale is usually carried out. However, such features which are extracted at a short term basis (several ms) have a great variability for the male and female speech and captures phoneme like characteristics which is not required. For the problem of gender identification, we actually need features that do not capture the linguistic information such as words or phonemes.

In this work we have made use of long term acoustic features called Piecewise Gaussian Modeling (PGM) features which we initially proposed for speech/music classification

(Harb and Chen, 2003b). The aim of PGM features for gender identification is twofold: (1) capturing long term spectral features that are discriminating between male's and female's voices, 2- modeling some aspects of the human perception of the sound, namely the short term memory aspect.

4.1.2.1. The Piecewise Gaussian Modeling (PGM). The PGM features are inspired by some aspects of the human perception and classification of sound signal. We propose that several features are essential for a human subject to classify a stimulus into audio classes, such as voice, music, applause, etc.

Consider the task of male/female classification by a human subject. For this task, classifying a short term audio segment, 10 to 20 ms, seems to be hard for human subjects. However, long term segments, 200 ms for instance, are easier to classify. Hence, we can state that long term features are essential for the human classification of speech by gender.

On the other hand, the insertion of short time segments of male speech, respectively female speech, into a long term time segment of female speech, respectively male speech, cannot be easily detected by human subjects. This supposes that human subjects classify audio segments by gender in a global manner. We hypothesise therefore that, human subjects when classifying a stimulus by gender base their decision on their actual auditory memory. That is, they classify a short term segment based on its correlation with the past auditory events.

For an offline mode, the model can be seen as a sliding window, the ITW window, on the signal where the Gaussian parameters are estimated, with a time stepΔITW. This is what we call the Piecewise Gaussian Modelling (PGM) of the signal. The time precision of the PGM modelling is equal to delta ITW (Typical values for multimedia applications are of the order of a few seconds). In this study, we use 1 second windows.

Formally, let $s(t)$ be the audio signal and t the time index. The short term spectral vectors, such as the Mel Frequency Spectral Vectors (MFSC), are: $\vec{X}_t, t = 1 \ldots N * T$ where N and T are two constants. T refers to the number of short term spectral vectors contained in a ITW window. For instance if every 10 ms one vector is obtained and the ITW window is 1 second, then $T = 100$. N refers to the number of ITW windows.

The PGM consists of modeling a set of T consecutive short term spectral vectors by one Gaussian model. That is, $N * T$ short term spectral vectors will be modeled by N

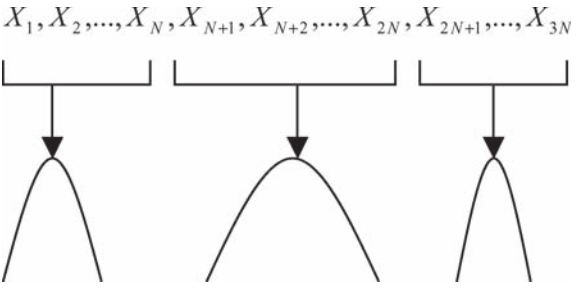

Figure 2. Piecewise Gaussian Modeling (PGM).

Gaussians.

$$\{\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_{N.T}\} \rightarrow \{\mathrm{M}_1(\vec{\mu}_1, \vec{\sigma}_1), \mathrm{M}_2(\vec{\mu}_2, \vec{\sigma}_2), \ldots, \mathrm{M}_N(\vec{\mu}_N, \vec{\sigma}_N)\}$$

With $\mathrm{M}_i(\vec{\mu}_i, \vec{\sigma}_i)$ is the i-th Gaussian expressed by its mean vector $\vec{\mu}_i$ and its variance vector $\vec{\sigma}_i$

$$\vec{\mu}_i = \frac{1}{T} \sum_{t=(i-1)N+1}^{iN} \vec{X}_t$$

$$\vec{\sigma}_i = \frac{1}{T} \sum_{t=(i-1)N+1}^{iN} (\vec{X}_t - \vec{\mu}_i) \cdot (\vec{X}_t - \vec{\mu}_i)^T$$

The PGM features are the concatenation of the mean and variance normalized by their respective maximum.

This modeling scheme has several advantages over the use of short term spectral vectors.

The PGM features do capture neither phoneme-like features nor word-like features. They capture the dynamics of the speech and the distribution of the energy in each frequency channel contained in a long term window which may better convey gender related attributes.

4.2. Complexity analysis

We studied the complexity of the male/female classification problem for each of the features presented above. Such a study can provide valuable information about the discrimination power of each feature for this classification problem.

We studied the short term spectral features and the PGM features. The complexity measures that we used are the Fisher's Discriminant Ratio, the Volume of Overlap Region, and the classification accuracy of a decision tree algorithm.

The data used for the complexity measures is a subset of the train_F dataset (see Section 6) containing 1000 seconds of male speech and 1000 seconds of female speech.

4.2.1. Fisher's Discriminant ratio. The Fisher's Discriminant ratio permits the estimation of the discrimination capability in each feature dimension. It is given by:

$$f(d) = \frac{(\mu_{1d} - \mu_{2d})^2}{\sigma_{1d}^2 + \sigma_{2d}^2}$$

Where $f(d)$ is the Fisher's Discriminant ratio for the feature dimension "d", and μ_{1d}, μ_{2d} σ_{1d}^2, σ_{2d}^2 are respectively the means and variances of classes "1"and "2" in the feature dimension "d".

As suggested in Ho and Basu (2002) we use the maximum of $f(d)$ over all dimensions. The higher the fisher discriminate ratio is the better are the features for the given classification problem.

Table 1. Fisher's discriminant ratio for 1000 s of male speech and 1000 s of female speech with PGM-features and short term spectral features.

	Short term spectral features	PGM-features
F	15.3	70.9

Clearly, as shown in Table 1, the PGM-features show higher Fisher's discriminant ratio than classic short term spectral features for this specific classification problem. These results suggest that PGM-features are more suitable than the short term spectral features for gender identification.

4.2.2. Volume of Overlap Region. The Volume of Overlap Region is another measure to analyze the complexity of a classification problem; it calculates the overlap between the classes in a selected feature space (Ho and Basu, 2002).

This can be measured by calculating the maximum and the minimum of the feature values in each feature dimension and then calculating the length of overlap for each dimension. The volume of overlap will be the product of the overlap lengths for all dimensions.

It is given by the following equation:

$$VOR = \prod \frac{MIN(\max(f_i, c_1), \max(f_i, c_2)) - MAX(\min(f_i, c_1), \min(f_i, c_2))}{MAX(\max(f_i, c_1)\max(f_i, c_2)) - MIN(\min(f_i, c_1), \min(f_i, c_2))}$$

Where $\max(f_i, c_1)$ and $\min(f_i, c_1)$ are respectively the maximum and the minimum values of the feature f_i for the class c_1 (resp. c_2).$i = 1, \ldots, d$ for a d-dimensional feature space.

The VOR is zero (if the overlap is negative then it is set to zero) if there is at least a feature dimension in which the two classes do not overlap.

With no contradiction to the Fisher's Discriminant ratio results presented in the previous section, the PGM-features show less overlap between male and female classes than short term spectral features as illustrated by Table 2. From this point of view PGM-features are better discriminator between males' and females' voices than the short term spectral features.

4.2.3. Decision trees. Decision trees are well known techniques in data mining domain. They can be used as classifiers as they aim at building rules in an IF THEN fashion permitting the decision about the class of a sample given its different attributes, or features.

We used the SIPINA one, a decision tree algorithm (Zighed, 1992). A decision tree was built to discriminate between the males and females classes given the short term spectral attributes and another tree was built given the PGM-features. The decision trees were used

Table 2. Volume of Overlap Region for 1000 s of male speech and 1000 s of female speech with PGM-features and short term spectral features.

	Short term spectral features	PGM-features
VOR	7.473 E-3	1.05673 E-17

Table 3. Error Rate of a Decision Tree trained on 1000 s of male speech and 1000 s of female speech and tested on the training data for PGM-features and short term spectral features.

	Short term spectral features	PGM-features
Error rate %	20	15.8

to classify the training data and the classification error rate was used as an evaluation of each of the features, namely short term spectral features and PGM-features.

As shown in the Table 3, the PGM-features perform clearly better for this classification problem hence motivating their use as an alternative of the classical short term spectral features for this problem of gender identification.

The study on the complexity measures suggests that the PGM-features are more discriminating than the short term spectral features. Moreover, the compactness of the features needed to describe the data using PGM-features motivates their use for gender identification. Thus, PGM-features have been retained as acoustic features beside the pitch related ones.

4.3. The classifier

The choice of a classifier for the gender detection problem in multimedia applications basically depends on the classification accuracy. Other considerations, such as the time needed for training and for the classification can also affect the choice since the volume of the data to be analyzed is huge. We investigated Gaussian Mixture Models (GMM), Multi Layer Perceptron (MLP), and Decision Tree classifiers. We do not include experiments about simple K-Nearest Neighbors classifiers since they are limited by low classification accuracy and high computation complexity.

4.3.1. Gaussian mixture models. Given a set of feature vectors, the GMM suppose that their probability distribution function is a combination of several Gaussians. Therefore GMM are a compact representation for a given classification problem since the information is embedded in the Gaussian parameters. GMM are also fast in both the classification and the training processes.

In our experiments the GMM were trained using the Expectation Maximization algorithm and initialized using the k-means clustering algorithm.

Since we use the highly correlated PGM features as the basic acoustic features we expect that the GMM performs poorly.

4.3.2. Multi Layer Perceptron. MLP imposes no hypothesis on the distribution of the feature vectors. It tries to find a decision boundary, almost arbitrary, which is optimal for the discrimination between the feature vectors. The main drawback for MLPs is that the training time can be very long. However, we assume that if the features are good discriminators between the classes and if their values are well normalized the training process will be fast enough.

Table 4. Classification accuracy for several classifiers trained on the same training data and tested on the same test data.

	GMM (14 Gaussians)	Decision Tree	MLP (80 Hidden neurons)
Accuracy %	72.5	73.4	82.2

In all our experiments we use a Multi Layer Perceptron with one hidden layer, 80 hidden neurons, and 2 output neurons. The Error Back-Propagation algorithm is used for the training.

The MLP classifiers are suitable for the PGM features since the MLPs are not strongly affected by the correlation of the input features as the statistical classifiers are.

4.3.3. Decision Trees. Decision Trees are well known in the data mining literature. They are special kind of trees where each node is a question on a parameter and the road from the root to one leaf is a rule permitting the classification of a feature vector, or a set of parameters.

All the classifiers were trained on 200 seconds of male speech and 200 seconds of female speech extracted from the train_F dataset and tested on the entire train_F dataset, that is 2200 seconds, Section 6. The classifiers were trained using the PGM-features. The classification results are shown in Table 4.

The classification accuracy of the MLP is noticeably higher than that of GMM and Decision Tree classifiers as shown in Table 4. Notice that, as it is expected the GMM classifiers perform poorly on highly correlated features while the MLP are more suitable. The PGM features also suppose a diagonal hypothesis on the covariance matrix which may contribute to the poor performance of the GMM.

GMM and MLP would probably provide similar performance if a decorrelated version of the features is used. Nevertheless, it seams that the MLP treats reasonably well the correlation in the PGM features and hence it was retained as the basic classifier for the gender identification problem.

4.4. Information fusion

It can be argued that the pitch and the PGM features do not capture the same characteristics of the signal. The experiments in the Section 6.2 support this assumption. We hypothesize for instance that the combination of the different features can improve the performance of a gender classifier. An in depth analysis, which is out of the scope of this paper due to space limits, of the correlation between PGM and pitch features would be useful in this case to enable an optimal combination of classifiers and features.

Several ways can be used for the combination of the features. Combining the features at the feature vector level, figure 3, is one generally used technique especially in speech recognition (Rabiner, 1993). An alternative of the combination at the feature level is the combination at the classifiers' output level (Kirchoff and Bilmes, 1999; Wu et al., 1998), figure 4. That is, for each feature type one classifier is used and the combination of the classifiers' outputs constitutes the features combined output.

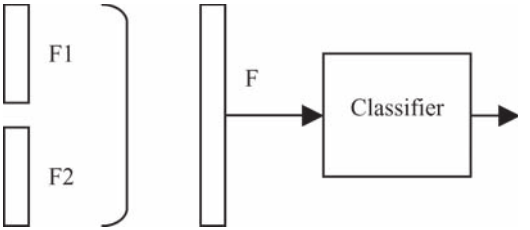

Figure 3. Fusion at the feature vector level.

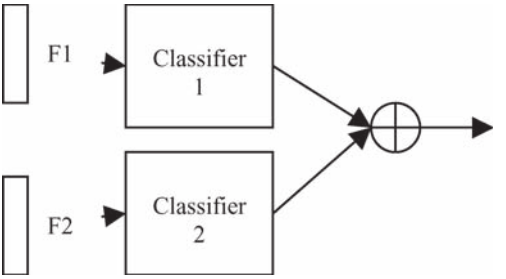

Figure 4. Fusion at the output level.

In this work we incorporated the combinations at the two levels, the features' level and the classifiers' output level. This combination as described in 7.1.2 showed improvements in the performance over one single combination technique.

5. Practical issues

Although the evaluation of a gender identifier is generally based on the frame classification accuracy, real world applications imply some improvements over the frame-based accuracy.

They are mainly motivated by the following two considerations.

5.1. Combining experts

The major drawback of the gradient descent training technique for Multi Layer Perceptron is the time needed for training when the number of training samples is large.

In the case of gender identification, the more acoustic conditions exist the more training data to describe such conditions is needed. Therefore, a general gender identification system with good accuracy must be trained on thousands of training samples making the scalability of the training process a critical issue.

In this work we investigated the splitting of the training data into manageable subsets of the training set, say 2000 training samples per subset. The subsets are obtained by sequentially segmenting the training data. One classifier, or expert, is trained on each subset using the

standard training algorithm. All experts trained on different subsets are then combined into one single expert. The combination is made by summing the outputs of the individual experts in order to obtain a single output layer for the resulting expert. Other functions such as multiplication or vote can also be used (Kittler and Alkoot 2003). However, in a previous work, our experimental results have shown that the performance obtained by multiplying or summing the outputs are similar while being particularly better than that of the majority vote (Harb et al., 2004). Therefore, the sum function is retained for the combination of the individual experts.

The output "j", male or female in this work, of the Mixture Expert "*ME*" is given by:

$$o_{\mathrm{ME}}^{(j)} = \sum_{i=1}^{N} \alpha_i . o_{Ei}^{(j)}$$

where "E_i" are the individual experts, "N" is the number of individual experts, and α_i is a weight associated to each individual expert. For the sake of simplicity the α_is are equal and they sum to one.

The obtained general expert is an expert trained on the complete training dataset while each individual expert is specialized on a subset of the data. Such a model can be trained on a complex classification problem since the complexity is treated by the mixture of the experts and each individual expert will be trained on a softer version of the classification problem. However, if the individual experts are trained on insufficient data, they possibly will become unstable and biased to the training data (Skurichina 1998).

A similar technique was applied for speech recognition (Mirghafori et al. 1994) applications with different experts trained on the speech from different speakers are trained.

5.2. Smoothing the classification results

The speech contained in audio-visual programs is continuous. This assumption leads us to incorporate the results of neighboring frames to smooth the classification results. We segment the speech signal using a metric-based approach and the KullBack-Leibler distance as measure of similarity between neighboring windows (Seigler et al., 1997). Each segment is assumed to contain one acoustic condition, in our case speech from the same gender. The labels of the frames in each segment are smoothed based on the average classification result for the entire segment. It is supposed that each segment contains speech from the same gender. However, we minimize the risk of mixing different genders in the same segment by decreasing the threshold used to segment the speech signal. Generally the classification results are slightly improved using such a smoothing technique, and in the worst cases the results do not change. In our system the threshold is set automatically to obtain a mean segments length of 3 seconds. That is, the local maxima in a sliding window of 3 seconds duration are selected as segment boundaries.

The used distance, the KullBack-Leibler (KL) distance, originates from the information theory (Cover and Thomas, 1991). It is a distance between two random variables. The original KL distance doesn't have the properties of a distance, but the symmetric KL is a distance. In the case of Gaussian distribution of the random variables the symmetric KL

distance is computed by:

$$KL2(X,Y) = \frac{\sigma_X^2}{\sigma_Y^2} + \frac{\sigma_Y^2}{\sigma_X^2} + (\mu_X - \mu_Y)^2\left(\frac{1}{\sigma_X^2} + \frac{1}{\sigma_Y^2}\right)$$

With $\sigma_X,\ \sigma_Y,\ \mu_X,\ \mu_Y$ are respectively the standard deviation of X and Y and the mean of X and Y.

In the case of audio segmentation, X and Y are the set of spectral vectors obtained from the ITW window X at time t, and the ITW window Y at time $t + T$ (T is the duration of the ITW window).

Notice that the variables used to compute the KL distance are the same features used in the PGM modeling; hence no new feature extraction is needed for the segmentation.

6. Experiments

Several experiments were carried out to evaluate the classifier for several classification conditions. We evaluated a "single Expert" and a "two Experts" gender identifiers with and without smoothing. The database used to evaluate the system consists of recordings from four French radio stations and one English radio station. Training data, Train_F, was extracted from the recording of a French radio station; it consists of speech from news programs and meetings. The data from other radio stations was used as the French test data, Test_F. Test_F data was also compressed with MPEG layer-3 coder at a 16 Kbps rate to obtain Test_F_mp3. Furthermore, 1600s were selected from the English radio station containing telephone speech, outdoor speech and studio speech constituting the Test_E dataset. A subset of the Switchboard dataset was also used for the evaluation. Table 5 shows the composition of the datasets used in the experiments. All the data was manually classified by gender. The amounts of time for male speech and female speech for each dataset were intentionally made equal.

6.1. System's architecture

An overview of our system for gender identification is presented in figure 5. It contains mainly two general modules: the feature extraction module and the classifier module.

Table 5. Evaluative dataset durations.

Dataset	duration (s)	Male speakers	Female speakers
Train_F	2200	12	7
Test_F	2000	10	7
Test_F_mp3	2000	10	7
Test_E	1600	9	7
Subset switchboard	4000	19	19

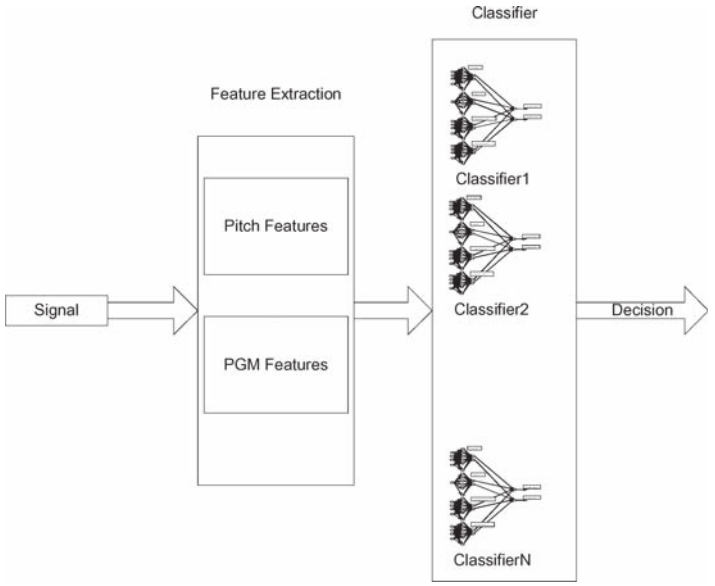

Figure 5. General overview of the gender identification system's architecture.

6.1.1. Feature extraction. The feature extraction module estimates pitch features and acoustic features.

As discussed earlier in Section 4.1., for each Integration Time Window (1 second in the experiments) the pitch and the PGM features are extracted in this module and constitute the input for the Classifier module. The PGM features are computed based on the Mel Frequency Spectral Coefficients (MFSC) which are obtained using Discrete Fourier Transform (DFT) with Hamming windows of 32 ms with 22 ms overlap and filtered using Mel scaled filterbank.

6.1.2. Classifier. The current system's classifier contains individual experts containing each 4 MLPs, figure 6. In each individual expert, one MLP uses the PGM features, one MLP uses the PGM-MinPitch features, one MLP uses the PGM-MeanPitch feature, and one MLP uses the PGM-AllPitch features.

Each MLP has 80 hidden neurons and 2 output neurons with output values ranging from 0 to 1 and corresponding to the probability of a feature vector to be female or male. The outputs of all the MLPs are summed to obtain two outputs for the classifier as it was presented in Section 5.1.

Each MLP is trained with no relation to the others using the Back Propagation algorithm. Once the MLPs are trained the individual expert is created and is already trained.

6.2. Classification accuracy for the single expert case

In this experiment the data form Train_F was used to train the system, and the data from Test_F was used to evaluate it. The system with one individual expert was used. The amount

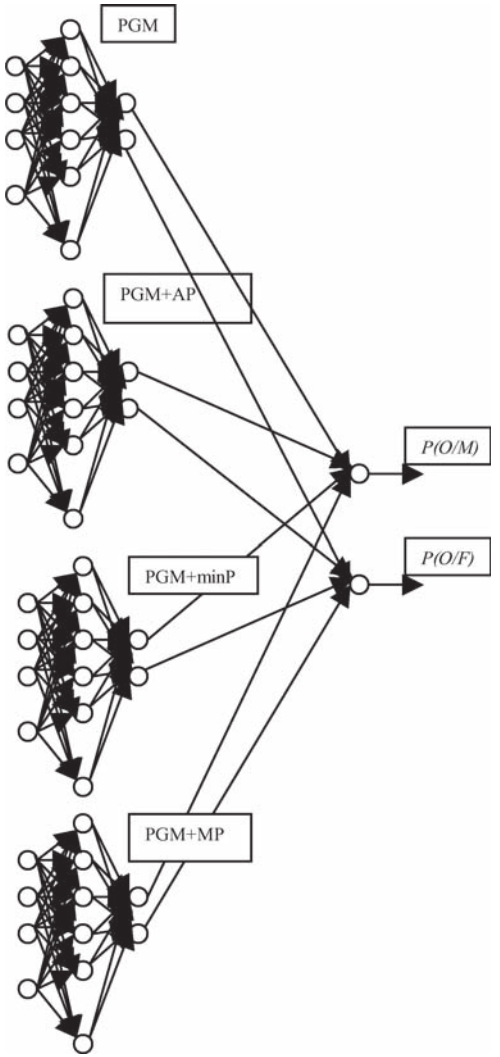

Figure 6. Architecture of the classifier used in the system.

of training data was changed in order to observe the effect of increasing training data on the classification accuracy. It is expected that by increasing the amount of training data the classification accuracy would increase.

The experimental results shown in figure 7 prove that indeed the classification accuracy increases by augmenting the size of the training data although the function is not monotonous. The accuracy of the classifier is about 75% when it is trained on 80 seconds of speech and it attains 90% when the classifier is trained on 2200 seconds of speech.

The overall classification results with a time precision of 1 second are 90.2% as shown in Table 6. These results are comparable to the reported results in the literature. A comparison

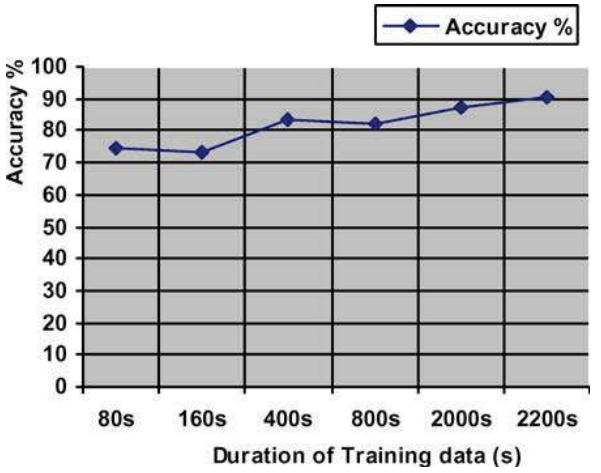

Figure 7. Classification accuracy rate as a function of the duration of the training data.

with the results reported in Tzanetakis and Cook (2002) (73%) shows the effectiveness of the proposed system over a straightforward general audio classifier approach based on MFCC features and Gaussian Mixture Models. Moreover, when the smoothing is performed, Section 6.2, the classification results are improved (92.7%).

Table 6 shows the results of each MLP contained in the expert and the overall accuracy of the expert. It is clear, that the expert performs better than its best MLP proving that the information fusion is motivated. Notice however, that the pitch information when included in the feature vector did not improve the performance. This may be due to the fact that the estimated pitch is not very reliable and that the pitch values overlap between male speech and female speech. Nevertheless, the overall accuracy can be improved, although slightly, when using pitch and acoustic information over the use of the PGM features alone.

We have carried out another experiment to observe the effectiveness of the proposed approach for language change and when the test data contains telephone and outdoor speech.

Table 6. Classification accuracy for radio data with and without smoothing.

	Male accuracy	Female accuracy	Total accuracy
Smoothing (3s mean segment length)	**95.2**	**90.1**	**92.7**
PGM	91.3	87.2	**89.3**
PGM-MinP	87.0	89.0	**88.0**
PGM-MP	88.6	86.3	**87.5**
PGM-AP	86.6	89.8	**88.2**
Average	88.4	88.1	**88.3**
Expert	**90.3**	**90.1**	**90.2**

Table 7. Classification accuracy for English data containing telephone and outdoor speech with and without smoothing.

	Male	Female	Average
Smoothing (3s mean segment length)	**93.2**	**84.2**	**88.7**
PGM	88.9	76.8	**82.9**
PGM-MinP	89.5	76.0	**82.8**
PGM-MP	88.5	79.9	**84.2**
PGM-AP	84.8	74.0	**79.4**
Average	87.9	76.7	**82.3**
Expert	**91.4**	**78.6**	**85.0**

The system with one individual expert was trained on 2200s from Train_F and tested on 1600s from Test_E data. The results shown in Table 7 demonstrate that the proposed approach is language independent though the performance is slightly degraded. Notice that the system was faced to language and channel changes.

We can also notice that in these two experiments, the feature vector which performs better is not the same. While MLP-PGM performs the best in Table 6, it is MLP-PGM-MP that performs the best in Table 7. However, in both cases the individual expert performs better than its best MLP. This is important since for different test conditions the best MLP is surpassed automatically by the individual expert.

The overall accuracy obtained by one single expert from the two datasets is thus 87%. This accuracy attains 90.7% when the smoothing is performed.

The last experiment for the individual expert case that we have carried out was to test the effectiveness of the gender identifier when the speech is compressed at low compression rates. The Test_F_mp3 was used as test set and Train_F was used for training. The classification accuracy for the individual expert did not degrade as compared to the accuracy obtained for the non-compressed data. This experiment tends to prove that the proposed system is robust to severe compression techniques as shown in Table 8

6.3. Classification accuracy for two experts case

In this experiment the system includes two individual experts each trained on half the training data, which is 1100 seconds of speech from Train_F dataset. The two experts are combined as described in Section 5.1. The system was tested on Test_F and Test_E datasets. The classification results are presented in Table 9.

Table 8. Classification accuracy of the individual expert when applied to compressed audio data.

	Male	Female	Total	Smoothing (3 s mean segment length)
Test_F_mp3	89.1	88.7	88.9	**90.0**

Table 9. Classification accuracy when two individual experts are used.

	Male	Female	Average	Average with Smoothing (3s mean segment length)
Test_F	88.5	89.6	89.1	**91.2**
Test_E	90.4	88.7	89.6	**94.6**
Total	89.5	89.2	89.4	**92.9**

As we can see, the classification results for the two datasets are improved from 87% for the single expert case to 89.4% when two individual experts are used instead. Moreover, in contrast to the case of one single expert, there is no bias toward any dataset or any gender when two individual experts are used. This is of extreme importance in multimedia applications since no information about the statistics of the data is given in advance. For instance one radio station may include English female speech only.

However, when the smoothing is performed the classification results attain an accuracy rate up to 93% on both datasets.

6.4. Classification accuracy for telephone speech

The speech material used in the previous experiments originates from noisy radio recordings, making the classification results look poor. The figure 8 shows a segment of speech from the Test_F database and one segment from the Switchboard database; we can see the nature of the data used in the previous experiments. It is important to evaluate the performance of the gender classifier using experimental conditions that are close to those used in the literature. Generally gender identification systems in the literature are evaluated on speech material obtained from the telephone data with a time precision larger than 5 seconds. Therefore, we used a subset of the switchboard database (Godfrey et al., 1992). 19 telephone conversations were used, 9 conversations for training the gender classifier and the remaining 10 conversations were used for testing the classifier. In average, the training data consists of 1000 seconds of male speech and 1000 seconds of female speech from 18 speakers,

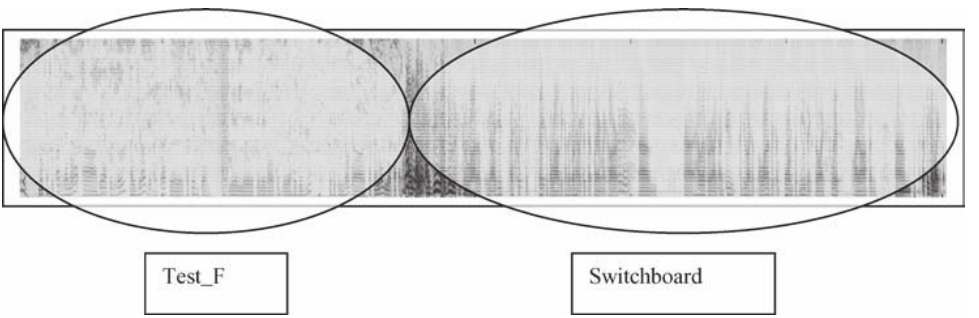

Figure 8. Example of the spectrogram of one speech segment from the Test_F database (on the left), and one speech segment from the Switchboard database (on the right).

Table 10. Classification accuracy for one individual expert on a subset of the Switchboard database.

	Male Accuracy	Female Accuracy	Average
1 s	88.2	98.9	93.5
5 s	97.0	100	98.5

9 male speakers and 9 female speakers. The test data corresponds to 1000 seconds of male speech and 1000 seconds of female speech from 10 male speakers and 10 female speakers different than those used for the training. One expert was used for this experiment and the results for 1 second and 5 seconds time precisions are presented in Table 10. As it is clearly shown in the table, the performance is particularly better than the performance for the case of the database used in the previous experiments. Gender identification for speech data obtained from multimedia sources seems to be harder than on speech data obtained from the telephone network. This is may be due to the existence of outdoor and noisy speech in the multimedia context. Moreover, as it is expected, by decreasing the time precision to 5 seconds, the classification accuracy increases significantly.

7. Conclusion

This paper presented a voice-based gender identification system using a general audio classifier. Several classifiers and features were studied. A combination of Piecewise Gaussian Modeling features and pitch-related features with a set of Neural Networks was shown to perform better than any individual classifier. The system was tested on adverse conditions of compression, channel mismatch and language change. It was shown how smoothing the classification results can improve the accuracy. When applied to telephone speech, the classification accuracy of the gender classifier was shown to be considerably better than when applied to speech from unrestricted radio sources.

Acknowledgments

This work has been partially supported by the Cyrano project within the French RNRT program. The first author was partially supported by Grant number 3691-2001 from the French ministry of research and technology. The authors would like to thank the anonymous reviewers whose critiques greatly improved the paper.

References

Acero, A. and Huang, X. (1996). Speaker and Gender Normalization for Continuous-Density Hidden Markov Models. In *Proc. 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-96*, vol. 1, 7–10, pp. 342 –345.

Cover Thomas M. and Thomas Joy A. (1991). Elements of Information Theory. In *Wiley Series in Telecommunications*. John Wiley and Sons.

DARPA TIMIT, Acoustic-Phonetic Continuous Speech Corpus, American National Institute of Standards and Technology, NTIS Order Number PB91–50565.

Delacourt, P. and Wellekens, C.J. (2000). DISTBIC: A Speaker-Based Segmentation for Audio Data Indexing. *Speech Communication*, 32, 111–126.

Godfrey, J.J., Holliman, E.C., and McDaniel, J. (1992). SWITCHBOARD Telephone Speech Corpus for Research and Development. In *Proc. IEEE ICASSP92 Conference*, pp. 517–520.

Harb, H. and Chen, L. (2003a). Gender Identification Using a General Audio Classifier. In *Proc. IEEE International Conference on Multimedia and Expo, ICME03*, 2, pp. 733–736.

Harb, H. and Chen, L. (2003b). Robust Speech Music Discrimination Using Spectrum's First Order Statistics and Neural Networks. In *Proc. IEEE International Symposium on Signal Processing and its Applications, ISSPA 2003*, pp. II-125–128.

Harb, H., Chen, L., and Auloge J.-Y. (2004). Mixture of Experts for Audio Classification: An Application to Male/Female Classification and Musical Genre Recognition. In *Proc. IEEE International Conference on Multimedia and Expo, ICME 2004.*

Haykin, S. (1994). *Neural Networks A Comprehensive Foundation*, Macmillan College Publishing Company.

Hemphill, C.T., Godfrey J.J., and Doddington, G.R. (1990). The ATIS Spoken Language Systems Pilot Corpus, In *DARPA Speech and Natural Language Workshop.*

Hess, W. (1983). *Pitch Determination of Speech Signals*. New York: Springer-Verlag.

Ho, T., and Basu, M. (2002). Complexity Measures of Supervised Classification Problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3), 289–300.

Huang, X.D., Lee, K.F., Hon, H.W., and Hwang, M.Y. (1991). Improved Acoustic Modeling with the SPHINX Speech Recognition System. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1991. ICASSP-91.,* Vol. 1, 14–17, pp. 345–348.

Jung, E., Schwarzbacher, A., and Lawlor, R. (2002). Implementation of Real-Time AMDF Pitch-Detection for Voice Gender Normalization. In *Proceedings of the 14th International Conference on Digital Signal Processing, 2002. DSP 2002*, Vol. 2, pp. 827–830.

Kirchoff, K. and Bilmes, J. (1999). Dynamic Classifier Combination in Hybrid Speech Recognition Systems Using Utterance-level Confidence Values. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. II-693–696.

Kittler, J. and Alkoot, F.M. (2003). Sum Versus Vote Fusion in Multiple Classifier Systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(1), 110–115.

Konig, Y. and Morgan, N. (1992). GDNN a Gender Dependent Neural Network for Continuous Speech Recognition. In *International Joint Conference on Neural Networks, 1992. IJCNN*, Vol. 2, 7–11, pp. 332–337.

Marston, D. (1998). Gender Adapted Speech Coding. In *Proc. 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1998. ICASSP '98*, Vol. 1, 12–15, pp. 357 –360.

Martland, P., Whiteside, S.P., Beet, S.W., and Baghai-Ravary, L. (1996). Analysis of Ten Vowel Sounds Across Gender and Regional Cultural Accent. In *Proc. Fourth International Conference on Spoken Language, 1996. ICSLP 96*, Vol. 4, 3–6, pp. 2231–2234.

Mirghafori, N., Morgan, N., and Bourlard, H. (1994). Parallel Training of MLP Probability Estimators for Speech Recognition a Gender-Based Approach, In *Proc. 1994 IEEE Workshop Neural Networks for Signal Processing* IV, 6–8, pp. 289–298.

Muthusama, Y.K., Cole, R.A., and Oshika, B.T. (1992). The OGI Multi-Language Telephone Speech Corpus. In *Proc. ICSLP 1992.*

Neti, C. and Roukos, S. (1997). Phone-Context Specific Gender-Dependent Acoustic-Models for Continuous Speech Recognition. In *Proc. 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 192–198.

Parris, E.S. and Carey, M. J. (1996). Language Independent Gender Identification In *Proc. IEEE ICASSP*, pp. 685–688.

Potamitis, I., Fakotakis, N., and Kokkinakis, G. (2002). Gender-Dependent and Speaker-Dependent Speech Enhancement. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002. (ICASSP '02),* Vol. 1, 13–17, pp. I-249–I-252.

Price, P., Fisher, W.M., Bernstein, J., and Pallett, D.S. (1988). The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 1988*, pp. I-291–294.

Rabiner, L.R. and Juang, B.H. (1993). *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall.

Rivarol, V., Farhat, A., and O'Shaughnessy, D. (1996). Robust Gender-Dependent Acoustic-Phonetic Modelling in Continuous Speech Recognition Based on a New Automatic Male Female Classification. In *Proc. Fourth International Conference on Spoken Language, 1996. ICSLP 96.,* Vol. 2, 3–6, pp. 1081–1084.

Ross, J.M. et al. (1974). Average Magnitude Difference Function Pitch Extractor. In *IEEE Transactions on Speech and Audio Processing*, 22, 353–362.

Seigler, M., Jain, U., Raj, B., and Stern, R. (1997). Automatic Segmentation, Classification, and Clustering of Broadcast News Audio. In *Proc. DARPA speech recognition workshop*.

Shimamura, T. and Kobayashi, H. (2001). Weighted Autocorrelation for Pitch Extraction of Noisy Speech. In *IEEE Transactions on Speech and Audio Processing*, 9(7), 727–730.

Skurichina, M. and Duin, R.P.W. (1998). Bagging for Linear Classifiers. *Pattern Recognition*, 31(7), 909–930.

Slomka, S., and Sridharan, S. (1997). Automatic Gender Identification Optimised For Language Independence. In *Proceeding of IEEE TENCON- Speech and Image Technologies for Computing and Telecommunications*, pp. 145–148.

Tzanetakis, G., and Cook, P. (2002). Musical Genre Classification of Audio Signals *IEEE Transactions on Speech and Audio Processing*, 10(5), 293–302.

Viswanathan, M., Beigi Homayoon S.M., and Tritschler, A. (2000). TranSegId: A system for Concurrent Speech Transcription, Speaker Segmentation and Speaker Identification. In *Proc. of the World Automation Congress, WAC2000*, Wailea, USA.

Wu, S., Kingsbury, B., Morgan, N., and Greenberg, S. (1998). Incorporating Information From Syllable-length Time Scales Into Automatic Speech Recognition. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing 1998*, pp. 721–724.

Zighed, D. A., Auray, J.P., and Duru, G. (1992). SIPINA: Méthode et logiciel (*in French*). Lacassagne.