Male and Female Voice Classification using GMMs

Gavan Keane - 11371985
Speech and Audio Processing - Assignment 2
30/12/2015

Abstract

The purpose of this report is to determine features that distinguish between male and female speech and to use these features to design a gender classification system. Here two features were tested, speaker fundamental frequency and MFCCs. Male and Female GMMs were trained using these features extracted from speech files taken from the TIMIT database. Fundamental frequency was shown to be a strong feature for the purposes of gender classification due to the high level of separability. The classification system based on fundamental frequency showed performance rates of over 95 % when used on clean speech signals. This performance however, dramatically fell off in the presence of even a small amount of noise. MFCCs proved to give a similarly high level of classification, however they also suffered from the same drop in performance when noise was introduced to the test utterances.

Contents

1. Introduction

A human listener is capable of inferring several characteristics of a person based purely on their speech such as their age, gender, ethnicity and emotional state. The most fundamental of these being gender identification. If gender can be correctly identified then it allows the use of gender dependent models when implementing more advanced speech processing systems such as speech and speaker recognition [9]. This greatly reduces the complexity of many speech processing problems and in many cases yields more successful results.

As in all classification systems the speech signal is first converted to a set of feature vectors. As with all speech based problems the choice of features is off some debate with no definably ideal feature available. The most common features used for gender classification through-out the current literature are Mel Frequency Cepstral Coefficients (MFCCs) [1] [2] [3] [5] [7] [9]. Less common is the choice of backend, i.e. the data structure used to cluster the feature vectors which is then used for classification. The most common choices of backend are GMMs [1] [9] and SVMs [2] [7]. Here we design a gender classifier using fundamental frequency and MFCCs as our features and GMMs as our backend. Feature vectors are extracted from a large sample set of male and female utterances taken from the TIMIT database. Separate GMMs are then fitted to the Male and Female features and classification is based on the likelihood of a new feature belonging to either one of the GMMs.

This report is structured as follows. Section 2 describes how both pitch and MFCCs were calculated from the TIMIT training data and how useful these features are likely to be for the problem of gender classification. This section also describes how a GMM model is fitted to the male and female feature vectors and how the MFCC and pitch based classifiers were designed. Finally this section suggests some alternative features which could be used in addition to give better performance. The results section demonstrates the performance of the MFCC and pitch based gender classifiers and discusses why each performed as it did. This section also discusses the effect of noise on the performance of the classifiers.

2. Classification Design

2.1. Test and Training Databases

The TIMIT database provides a large number (4620) of sample speech waveforms for 8 different dialects of North America consisting of a vast amount of speakers. All utterances are recorded at 16000 Hz and last between 5 and 12 seconds. The waveforms have a high SNR and the database contains roughly equal amounts of male and female speakers. Both the training and test data used throughout this report was taken from the TIMIT data base. Special care was taken to ensure that there was no overlap between the test and training data. The training data consisted of two separate databases, one containing only male speakers and one containing only female speakers. Both databases consisted of 300 utterances with 50 utterances taken from each of the first six

dialect regions. The test data consisted of 450 utterances taken from both male and female speakers from all 8 dialect regions (as second set of test data consisting of 150 male utterances and 150 female utterances was also created). As mentioned above none of the training data was repeated in the test data to ensure unbiased test conditions. For a more rigorous trial, the classifier should also be tested using utterances from a database other than TIMIT or from real world scenarios. Additionally 16 small databases, each consisting of ten utterances from male or female speakers from all eight dialect regions were used for small scale testing.

	Utterances	Dialect Regions	Male or Female Speakers
Male Training Data	300	1-6	Male
Female Training Data	300	1-6	Female
Test Data 1	450	1-8	Male and Female
Test Data 2	300	1-5	Male and Female

Table 1: Details of the training and test databases taken from TIMIT

2.2. Separation of Male and Female voices based on Fundamental Frequency

The most prominent difference between male and female speech is the fundamental frequency of the waveform. Studies on the topic [1] have shown that the majority of male speech has a fundamental frequency below 160 Hz while the majority of female speech has a fundamental frequency above 160 Hz. This suggests that it should be easy to separate male and female utterances based purely on the speakers pitch. [4] Proposes a gender recognition model based purely on speaker fundamental frequency while [1] and [5] propose methods which combine fundamental frequency and MFCCs to determine speaker gender.

The YIN pitch estimation algorithm [10] was used to determine the average pitch. As with all pitch estimators, pitch can only be determined from voiced speech. Similarly there are often instances where the estimated pitch is an octave above or below its true values. In order to ensure a high level of accuracy only clearly voiced sections of speech were used to determine fundamental frequency (these are the sections defined with the 'best' probability by the YIN algorithm). Additionally the following changes were made to the default YIN parameters for male and female speakers. Note that when training the gender classifier we know whether a given utterance is male or female. In the case of testing the pitch based gender classifier we don't initially know the gender of the utterance, thus different tuning parameters for the YIN algorithm must be used. Table 2 shows the parameters used for these three cases.

	Minimum frequency	Maximum Frequency	Window Length (samples)	Window Shift (samples)	Low Pass Filter
Male Speaker	60 Hz	200 Hz	400	100	900 Hz
Female Speaker	130 Hz	330 Hz	200	50	1100 Hz
Unknown Gender	60 Hz	330 Hz	300	75	1100 Hz

Table 2: Tuning Parameters used for the YIN algorithm for when the speaker is known to be male, when the speaker is known to be female and when the speaker is of unknown gender.

We only require a single value for pitch for each speaker. In general the fundamental frequency of a speaker doesn't change much, thus the average pitch of all the clearly voiced windows (see above) within each utterance is used for this single value. Taking the average also helps smooth out the effects of any incorrect pitch estimations (which are few given the above tuning parameters). Note that the median pitch can also be taken instead of the mean however this doesn't have much effect on the accuracy of the classifier (see table 6).

For each of the 300 male utterances and 300 female utterances from the training database the average pitch was determined using the above method. The histogram in figure 1 shows the average pitch values for all 600 utterances (male utterances are red and female utterances are blue)

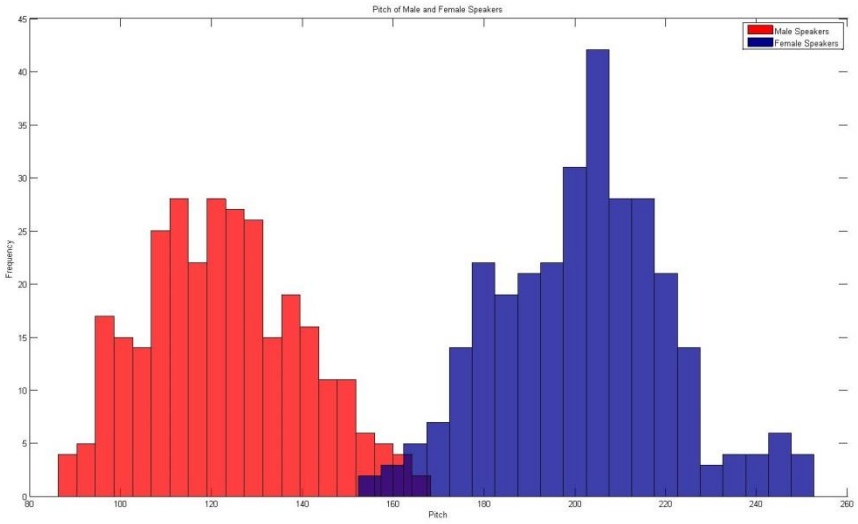


Figure 1: Histogram showing the average pitch for 300 male utterances (red) and 300 female utterances (blue)

From figure 1 it's easy to see that for the majority of cases male and female speakers can be separated based purely on gender. This suggests that pitch is a strong feature to use in a speech based gender classifier. A Gaussian mixture model was fitted to each the male, and female fundamental frequency datasets. Figure 2 shows the PDF for both the empirical data and the fitted GMMs.

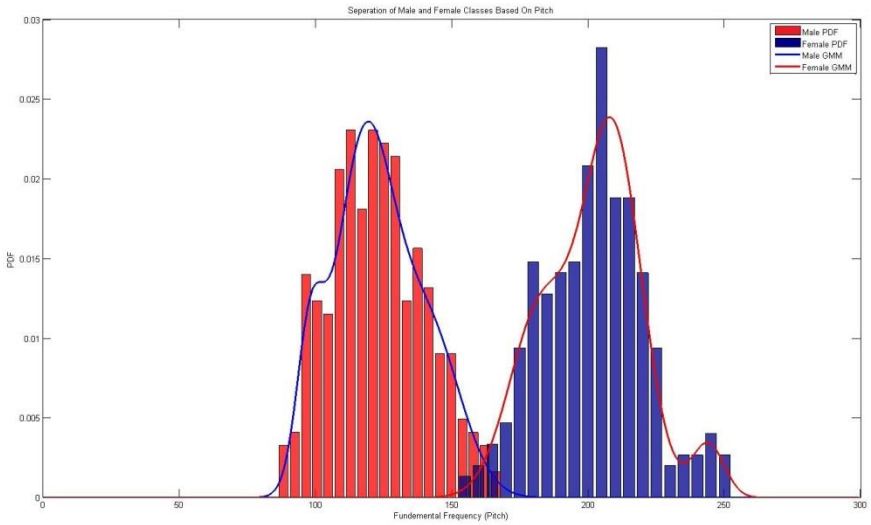


Figure 2: Bar charts show the PDF of the empirical pitch data (red for male and blue for female). Plots show the PDF of the fitted GMM distributions (blue for male and red for female)

The AIC algorithm (described in section 2.4) was used to determine the best number of components (mixtures) for each of the fitted GMMs. It was found that both the male and female distributions were best fitted by a GMM with 3 mixtures (the parameters of the three mixtures are shown in table 3). The important point to note from figure 2 is the cross over point (point at which female PDF is greater than male PDF). This occurs at 162 Hz (This agrees with the threshold proposed in [1]). A classifier based on this data will essentially be a thresholding operation where utterances with an average bitch below 162 Hz will be classed as male and utterances with an average pitch above 162 Hz will be classed as female. We can determine the percentage of correctly classed males and females within the training data from the CDF plots shown in Figure 3.

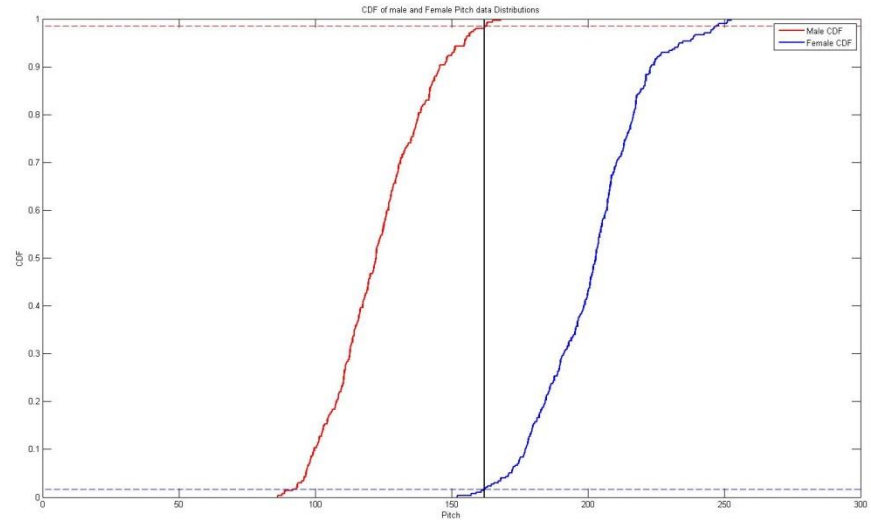


Figure 3: Shows the empirical CDFs of the male (red) and female (blue) average pitch distributions.

The thick black line is the threshold value determined from the GMM distributions. The dotted red line represents the probability of correctly classifying a male utterance from the training data as male and the dotted blue line represents the probability of incorrectly classifying female utterance from the training data as male. These percentages are shown in table 4. Note that while pitch does give a good degree of separability between the two genders it isn't difficult for a person to alter the pitch of their voice meaning that this system is easily fooled.

	Mixture 1		**Mixture 2**		**Mixture 3**	
	Mean	Variance	Mean	Variance	Mean	Variance
Male GMM	97.99	24.49	141.34	139.93	117.88	112.69
Female GMM	182.56	131.85	243.96	34.25	208.83	113.67
	Mixture 1		**Mixture 2**		**Mixture 3**	
	Weights		Weights		Weights	
Male GMM	0.1085		0.3073		0.5842	
Female GMM	0.3380		0.0495		0.6125	

Table 3: GMM distributions for male and female pitch with each using 300 utterances taken from the TIMIT database (all values are in Hz).

Male Classified as Male (%)	98.52 %
Male Classified as Female (%)	1.48 %
Female Classified as Female (%)	98.38 %
Female Classified as Male (%)	1.62 %

Table 4: Estimated performance of the Pitch based gender classifier based on the cumulative distributions of the training data

2.3. Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCCs) are a set of features commonly used in automated speech recognition. MFCCs were introduced by Davis and Mermelstein in the 1980s and have been state of the art ever since [8]. Speech generated by a human is filtered by the vocal tract and the shape of the vocal tract determines what sound comes out. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum of the speech waveform. MFCCs are used to accurately represent this envelope [8]. Papers [7] and [9] both propose speech based gender classifiers were MFCCs are the only features. Paper [7] models the features using a GMM while [9] models the features using a SVM. Both papers report correct classification rates over 90%. The literature seems to suggest that MFCCs should be a good feature for classifying gender.

The process for calculating the MFCCs of an utterance is as follows

1) The speech waveform is segmented into short frames. Usually each frame lasts 25 ms and there is an overlap of 10 ms between frames [8]. This is due to the signal being relatively stationary over this time period. Short frames don't contain enough samples for a reliable estimate whereas the signal will start to change over longer frame durations

2) Each frame is multiplied by a hamming window and an estimate of the periodogram of the power spectrum is calculated. The periodogram identifies which frequencies are present in each frame.
3) A Mel frequency filterbank (Figure 4) is calculated and each filter is multiplied by each frame to give a coefficient for that frame. Generally the Mel Filterbank has 26 filters so each frame will have 26 coefficients. The first filter is very narrow and gives an indication of how much energy exists near 0 Hz.

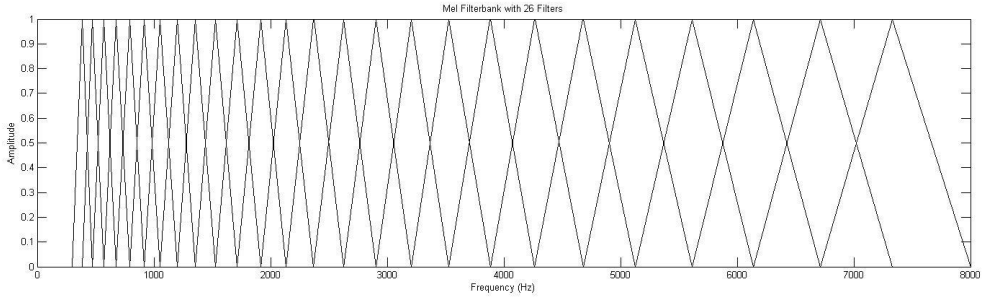


Figure 4: Mel Filterbank from 300Hz to 8000Hz containing 26 filters

4) The logarithm of the filterbank energies is taken for each frame. This allows us to use cepstral mean subtraction.
5) Finally the MFCC coefficients are obtained by taking the DCT of the log filterbank energies. This needs to be done since the filterbanks are overlapping, thus they are heavily correlated. By decorrelating the filterbank energies we can use diagonal covariance matrices to model the features in our GMM [8].
6) It is not necessary to keep all the MFCCs for each frame. Paper [8] suggests that keeping the higher DCT coefficients may degrade the performance of the classifier. Additionally [9] suggests that the best number of MFCCs to use with a GMM is 12.

In order to ensure correct MFCC calculations two separate methods were used to calculate the coefficients. The first method uses the MFCC function from the voicebox toolbox [11]. The second method uses a function written by Kamil Wojcicki for mathworks. The following table shows the tuning parameters used to calculate the MFCCs for both functions.

Sampling Rate	Window Length	Window Overlap	Window method	Frequency Range	Number of filters	Liftering Parameter
16 kHz	25 ms	10 ms	Hamming	300 – 8000 Hz	26	22

Table 5: Tuning parameters used to calculate the MFCCs for each utterance

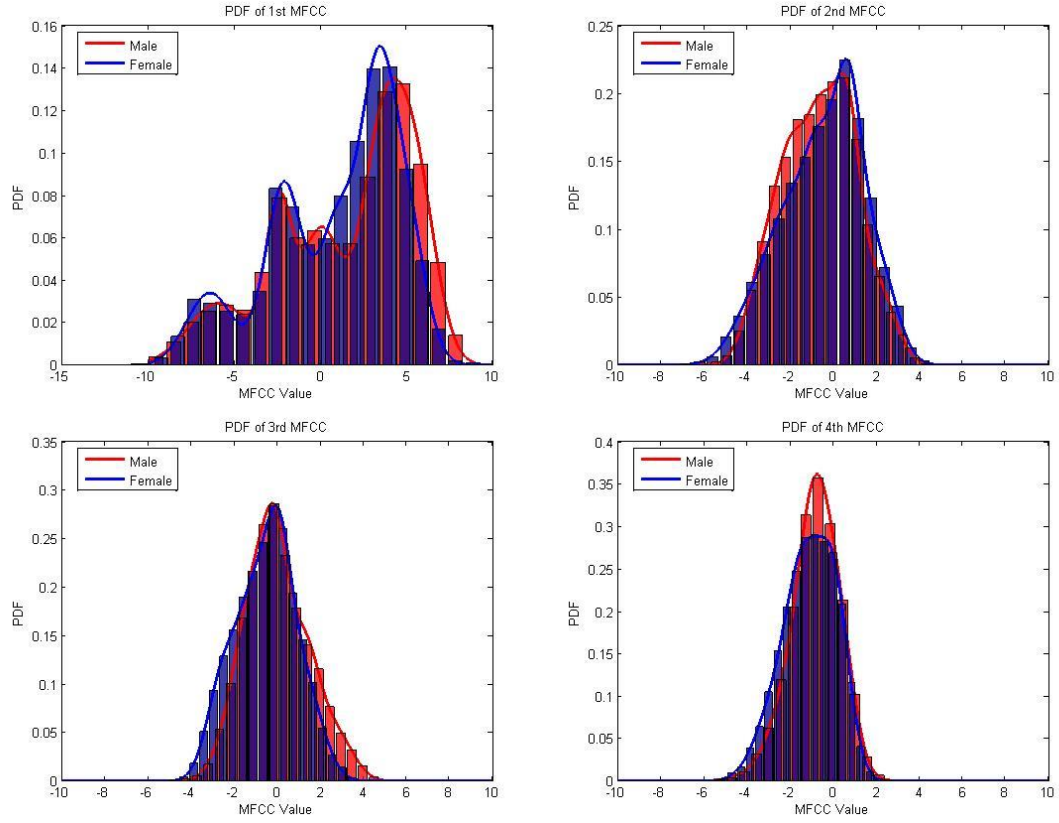
12 MFCC coefficients were used when designing our gender classifier. This is discussed in more detail in the results section. As well as the standard MFCCs the delta and delta-delta coefficients can also be calculated (would add 24 additional features per frame). The delta coefficient are calculated using the following formula where N is generally equal to 2, C_n are the static coefficients (delta coefficients if calculating delta-deltas) and t is the frame index

$$d_t = \frac{\sum_{n=1}^{N} n\left(C_{t,+n} - C_{t,-n}\right)}{2\sum_{b=1}^{N} n^2}$$

Paper [5] suggests that the inclusion of deltas and delta-deltas doesn't improve recognition and that they will only serve to increase complexity and computation time.

Despite the promise of MFCCs from many of the papers, MFCCs appear to be a very poor feature for preforming gender identification. The reason being that MFCCs calculated for male and female utterances (within the TIMIT database) have nearly identical distributions. Figure 5 shows overlapping histograms of the first 6 MFCCs (univariate case) for male (red) and female (blue) utterances. Additionally Figure 6 shows the GMM distributions fitted to the first four pairs (bivariate case) of MFCC coefficients (i.e. 1 and 2, 3 and 4). Again we see that the male and female distributions almost entirely overlap. Higher dimensional cases can't be observed visually however, the trend of inseparability likely continues.

Since the male and female MFCCs are almost entirely inseparable any classifier built from this data would have likely have a classification rate only slightly better than random guessing (slightly over 50 % accurate). This idea of inseparability is supported by [18] which also found a large deal of overlap between male and female MFCCs when using a GMM. Despite this, papers [7] and [9] claim a gender classification rate of over 90% when using just MFCCs as their classification features. Paper [9] claims a classification rate of 100% when using 24 MFCCs and a 12 component mixture model.



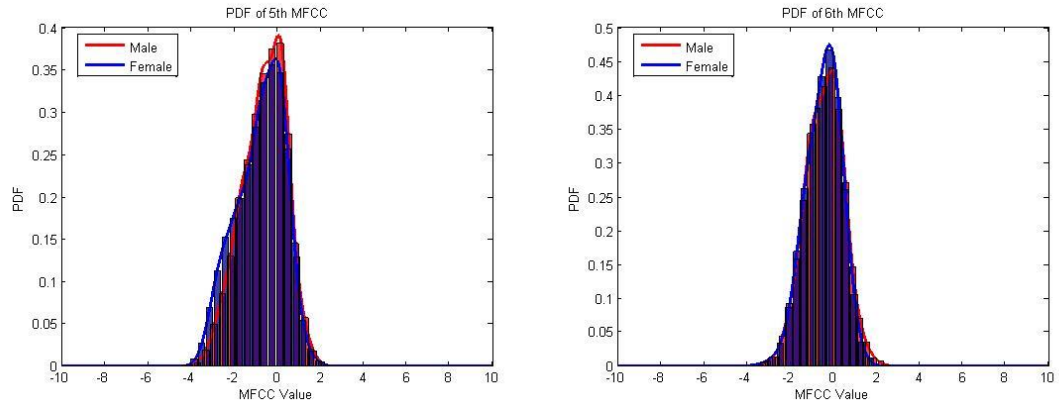


Figure 5: Plots showing the distribution of the first 6 MFCC coefficients calculated using 50 male utterance (red) and 50 female utterances (blue). The line plots are GMM models fitted to each distribution

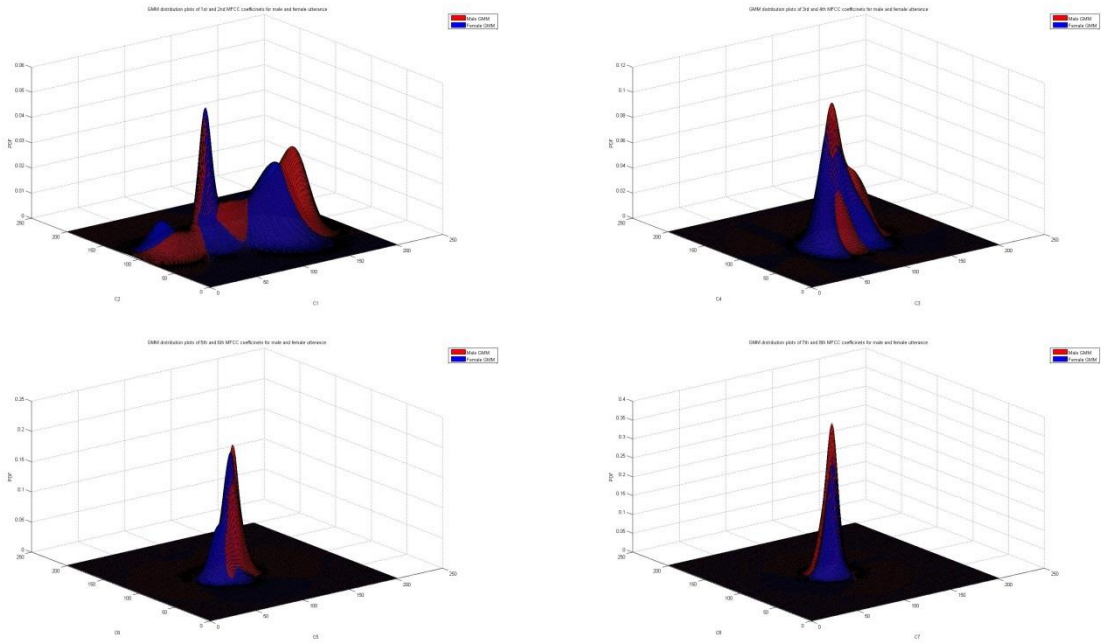


Figure 6: Plots showing the GMM distributions fitted to bivariate MFCC coefficients for 50 male (red) and 50 female (blue) utterances. Note that diagonal covariance matrices were used for the MFCCs (see above)

2.4. Gaussian Mixture Models

In this report a Gaussian mixture model is used to model the distribution of feature vectors for male and female utterances. Classification is decided based on the likelihood that a set of feature vectors belongs to either the male or female GMM. The sample is classified based on which GMM has the higher likelihood.

A Gaussian mixture model is a parametric probability density function represented as a weighted sum of Gaussian component densities. An example of a univariate GMM is shown in table 3. Given a set of training feature vectors, GMM parameters are estimated using either the expectation

maximisation algorithm or the maximum a posteriori estimate (this requires a well-trained prior). Here the EM algorithm is used to estimate the GMM parameters. A multivariate GMM is described by the following equations where w_i are the component weights, μ_i are the component means and Σ_i are the component covariance matrices (diagonal matrices in our case since the DCT decorrelates the MFCC coefficients). M is the number of components and λ is the GMM model.

$$P(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} w_i g(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma_i})$$

$$g(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma_i}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma_i^{-1}}(\boldsymbol{x} - \boldsymbol{\mu}_i)\right\}$$

The idea behind the expectation maximisation algorithm is to find model parameters which maximise the likelihood of the GMM given a sequence of training vectors $\overline{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$. Assuming independence between the vectors this can be written as [6]

$$P(\overline{X}|\lambda) = \prod_{t=1}^{T} p(\boldsymbol{x}_t|\lambda)$$

Unfortunately this is a non-linear function of λ so direct maximisation isn't possible however, an estimation of the maximum likelihood can be obtained iteratively. This is done by starting with an initial model λ and then estimating a new model $\overline{\lambda}$ such that $P(\overline{X}|\overline{\lambda}) \geq P(\overline{X}|\lambda)$. This new model then becomes the initial model in the next iteration until some convergence threshold is reached. In our case the maximum number of iterations allowed was limited to 1,000 and the original initial model was derived using K-nearest neighbour clustering. On each iteration, the following three formulas are used which guarantee a monotonic increase in the model's likelihood value [6]

$$\overline{w}_i = \frac{1}{T}\sum_{t=1}^{T} P(i|\boldsymbol{x}_t, \lambda)$$

$$\overline{\mu}_i = \frac{\sum_{t=1}^{T} P(i|\boldsymbol{x}_t, \lambda)\boldsymbol{x}_t}{\sum_{t=1}^{T} P(i|\boldsymbol{x}_t, \lambda)}$$

$$\overline{\sigma}_i^2 \frac{\sum_{t=1}^{T} P(i|\boldsymbol{x}_t, \lambda)\boldsymbol{x}_t^2}{\sum_{t=1}^{T} P(i|\boldsymbol{x}_t, \lambda)} - \overline{\mu}_i$$

One issue when using Gaussian mixture models is that it can be difficult to know the best number of components to use. The Akaike information criterion (AIC) is a measure of the relative quality of statistical models for a given set of data and can be used to select a GMM which contains the most information about the observed data while penalising models which are over-parameterised [13]. That is, the AIC method balances the risk due to bias when a low dimensional model is selected and

the risk due to an increase in variance when a high dimensional model is selected. The AIC value for a model is calculated as

$$AIC = 2k - 2\ln(L)$$

Here L is the maximum likelihood of the modal and k is the number of model parameters. Our goal is to minimise this value to determine the best fit GMM. This is done by fitting GMMs to our training feature vectors with one to ten components (mixtures) and then selecting the GMM which produces the lowest AIC value as the final model. In the results section we discuss whether the model chosen by the AIC algorithm really gives the best classification results.

2.5. Classification using Pitch

As demonstrated in section 2.2 male and female speech can be correctly classified based purely on pitch in over 95 % of cases (based on observations discussed in section 2.2). Here we outline the framework for a gender classifier based purely on pitch. The results and performance of this pitch based classifier are covered in section 3.1. Figure 7 shows a block diagram demonstrating the design process of our pitch based gender classifier. The classifier was designed using the following steps

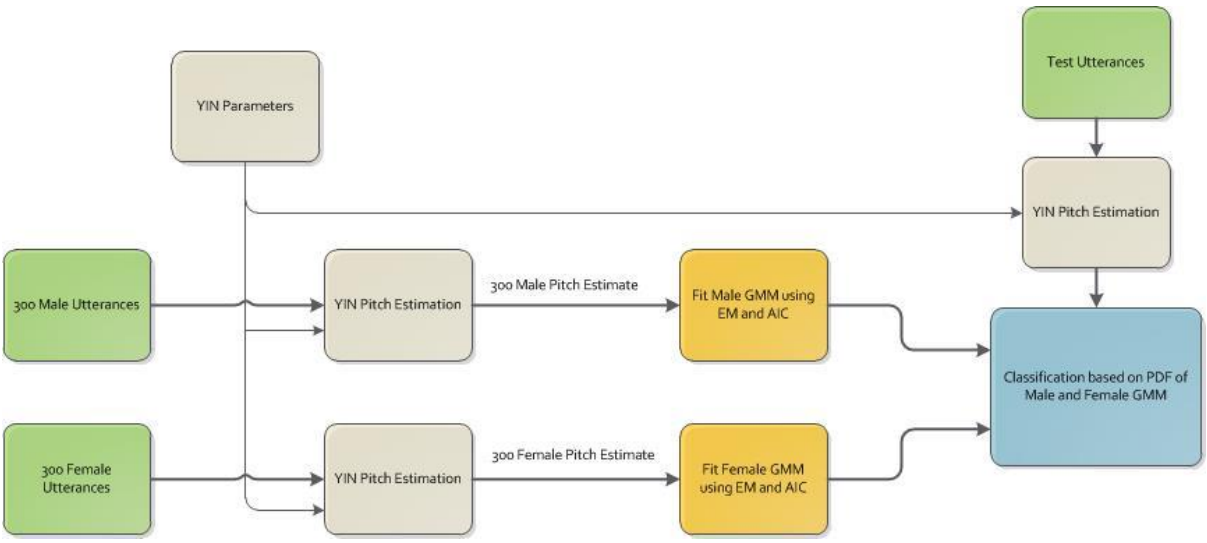


Figure 7: Block Diagram showing design of pitch based gender classifier

1) An average pitch estimate was obtained for each of the 300 male utterances and 300 female utterances contained within our training data set (see section 2.1). The pitch estimate is obtained using the YIN algorithm as discussed in section 2.2. Only sections of clearly voiced speech were used when estimating the pitch, additionally any estimates outside the range of 60 Hz to 350 Hz were removed as they're clearly incorrect. The final pitch estimate is taken as the average of the remaining windowed estimates, the idea being that inaccuracies in the estimate will go to zero due to the law of large numbers. As a comparison the median value was also used however this provided no noticeable difference in the rate of classification.
2) A univariate GMM model was fitted to each the male pitch estimates and the female pitch estimates as described in section 2.4. GMMs using 1 to 10 components were fitted to the

training data. For both the male and female data it was found that 3 components best fit the data. Table 3 shows the GMM models that were fitted to the male and female data.

3) Classification was done by estimating the pitch for each of the 450 test utterances in the same manner as describe d in step one. The probability density function was then calculated at this value of pitch for both the male and female GMM. The utterance was classified to which ever gender had the larger value.

2.6. Classification Using MFCCs

Despite the suggestion in section 2.3 that MFCCs would be a poor feature for determining gender a classifier was designed using MFCCs as the only features. This was motivated by two papers [7] [9] which claimed a high classification rate (over 90 %) for gender classification using MFCCs and GMMs. The results and performance of our MFCC based gender classifier are shown in section 3.2. Figure 8 shows a block diagram outlining the design of our classifier. The classifier was designed using the following steps

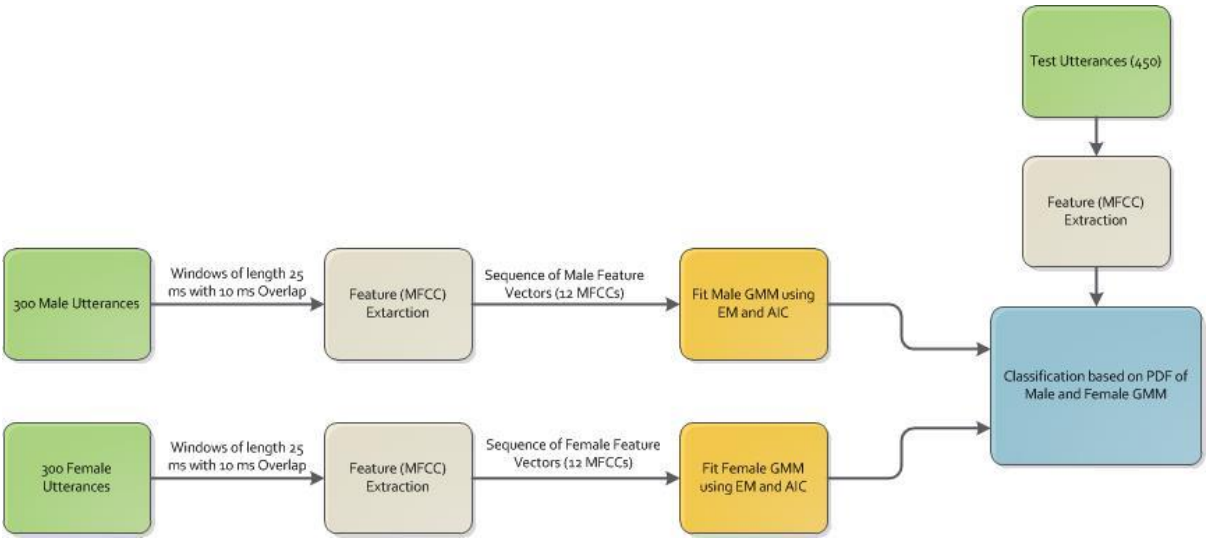


Figure 8: Block Diagram showing design of MFCC based gender classifier

1) Each of the 300 male utterances and 300 female utterances from our training data were windowed into sections of length 25 ms (400 samples) with an overlap of 10 ms (160 samples). MFCCs were calculated for each window in each utterance as described in section 2.3 using a Mel filterbank with 26 filters. Classifiers were designed using the first 6, first 12 and first 24 MFCCs to compare the rate of classification (see section 3.2). The delta and delta-delta coefficients were not used as it has been shown they don't improve the rate of gender classification [5].
2) A Multivariate GMM model was fitted to each the male MFCCs and the female MFCCs as described in section 2.4. GMMs using 1 to 10 components were fitted to the training data and the AIC value was used to determine the best fitting GMM. The EM algorithm was initialised using K-nearest neighbour clustering and the covariance matrices for each of the components were set to be diagonal since the DCT decorrelates the MFCCs

3) Classification was done by calculating the MFCCs for each of the 450 test utterances in the same manner as step 1. The PDF was then calculated at the MFCC feature vectors for each windowed section in the utterance for both the male and female GMM. The result of this was PDF values for each windowed section of the utterance showing the likelihood that it was male or female. 3 approaches were taken to classifying the utterance based on these PDF values. The first approach was to take the average value of all the male PDF values and compare it to the average value of all the female PDF values. The utterance would then be classified as which gender had the larger average value. The second method is the same as the first except the median value was used instead of the average. The third method was to compare the PDF values for each window (section). Classification of the utterance was then determined as which gender the majority of the windowed sections were classified as. As it turns out this is the only method that produces accurate classification.

2.7. Possible Alternative Features

The main distinguishing feature between male and female speech is the pitch however, simply increasing the pitch of a male speaker doesn't make them sound like a female speaker nor vice-versa. This tells us that there are other possible features which could be used to determine the gender of a speaker. Possible examples of these features are timbre, rhythm, stress and tempo. While paper [15] proposes a method for determining speaker rhythm and paper [16] proposes a method for determining speaker timbre none of these features are particularly easy to determine accurately nor is there any real suggestion that they would greatly improve a gender classifier.

Another possible approach to gender classification is based on high level features such as the number of pauses, number of paralinguistic features (umm and yeah) and number of subject changes. It has been shown that these sorts of features can distinguish between male and female speakers [17], they are however, likely to be highly situational biased.

Finally both pitch and MFCCs have shown to be sensitive to noise (see results section). A possible alternative feature would be Power Normalised Spectral Coefficients [19]. These are recently discovered features which can perform similarly to MFCCs but are more robust to noise.

3. Results

3.1. Classification Using Pitch

As mentioned in section 2.5 the estimated pitch for each utterance was calculated as the average of the YIN pitch estimates for each window of clearly voiced speech. As a comparison the pitch estimate of the utterance was also taken as the median value of the pitch estimates for each window. The effects this had on classification for both the training and test data are shown in table 6. Table 6 also shows the pitch based gender classifiers performance when tested using the training data and when using the test data.

	Correct Male	Correct Female	Incorrect Male	Incorrect Female	Overall Correct %
Training Data (Median)	292	288	8	12	96.67 %
Training Data (Mean)	294	281	6	19	95.83 %
Test Data (Median)	336	105	7	0	98.44 %
Test Data (Mean)	332	105	13	0	97.11 %

Table 6: Comparison of pitch based gender classifier performance when tested using training data and test data. Also shows comparison between median pitch estimation and mean pitch estimation for each utterance

There are two important points to take away from table 6. Firstly, and rather surprisingly the test data has a better rate of classification than the training data. This suggests that either the training data or the test data isn't a good representative set. One possible reason for this occurring is that the training data contains an even amount of male and female utterances whereas the test data contains 3.5 times as many male speakers as female speakers. If the classifier is biased toward male speakers then this would explain the higher classification percentage when there are more male utterances. An argument against this however is that the classifier is 100% accurate with female utterances taken from the test data, this would suggest that the classifier is biased towards female utterances.

	Correct Male	Correct Female	Incorrect Male	Incorrect Female	Overall Correct %
Test Data (revised)	150	148	0	2	99.3 %

Table 7: shows classification performance of pitch based gender classifier with revised test utterances

In order to confirm these results the classifier was retested with 300 new test samples (150 male and 150 female) again, taken from the TIMIT database. The results are shown in table 7. Again there is the issue that the test data is scoring higher than the training data. In this case however it is the female utterances that are being incorrectly classified. Another interesting point revealed by table 6 is that the classification performance of the training data is less than the predicted values in table 4, particularly for the female utterances. The final point of interest is that for the test data the median approach preforms better while the mean approach preforms better for the training data. This seems to suggest that using the mean method or the median method doesn't have a huge amount of influence on whether one method preforms better than the other.

As mentioned in section 2.4 it is difficult to know in advance how many components are required to best fit a GMM to some training data. The performance metrics shown in tables 6 and 7 were obtained using 3 components for both the male and female GMM (these being the optimal according to the AIC values). Figure 9 shows a plot of the classification performance as the number of components in the male and female GMMs is increased from 1 to 10. From this plot it is clear that the effect of the number of components on classification performance is minimal. One point of interest is that the best performance is provided when there are five components and not three as

suggested by the AIC algorithm. This suggests that the best fitted model doesn't necessarily result in the highest rate of classification. The other point to note is that performance doesn't continue to increase with an increased numbers of components. This suggests that over parameterising the model will result in a small drop in performance.

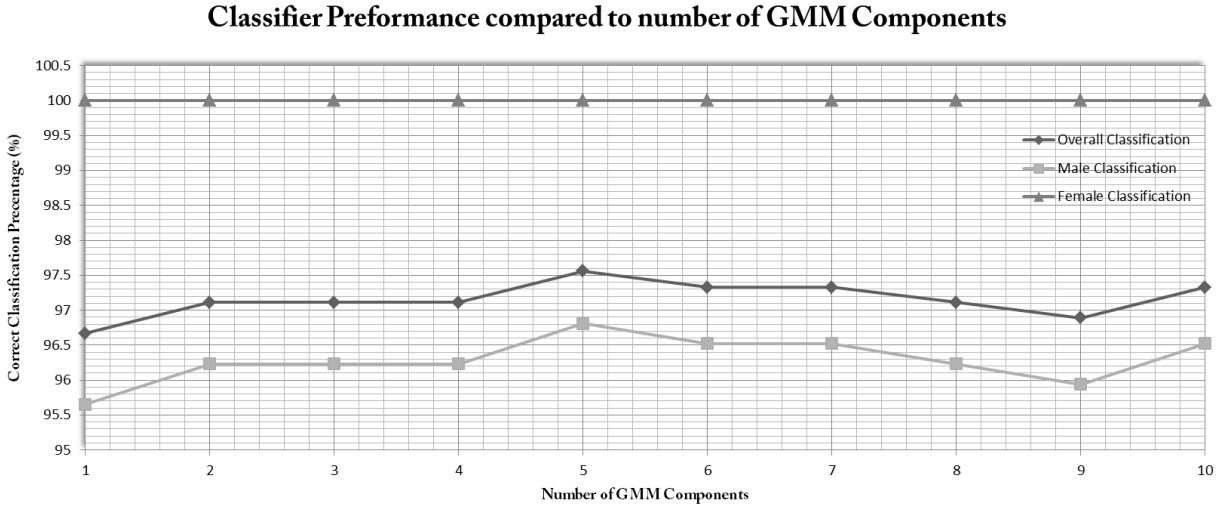


Figure 9: Graph showing the effect the number of GMM components has on classification

The tables and figures above show that pitch based gender classification can achieve a very high performance level operating on noiseless speech samples. The primary difficulty with pitch based gender classification is estimating a pitch value of the utterance. This becomes increasingly difficult in the presence of noise. To demonstrate this, white Gaussian noise was added to each of the test utterances before estimating the pitch. Figure 10 shows the effect of noise on classification performance with noise signals ranging from 5 dB to 60 dB.

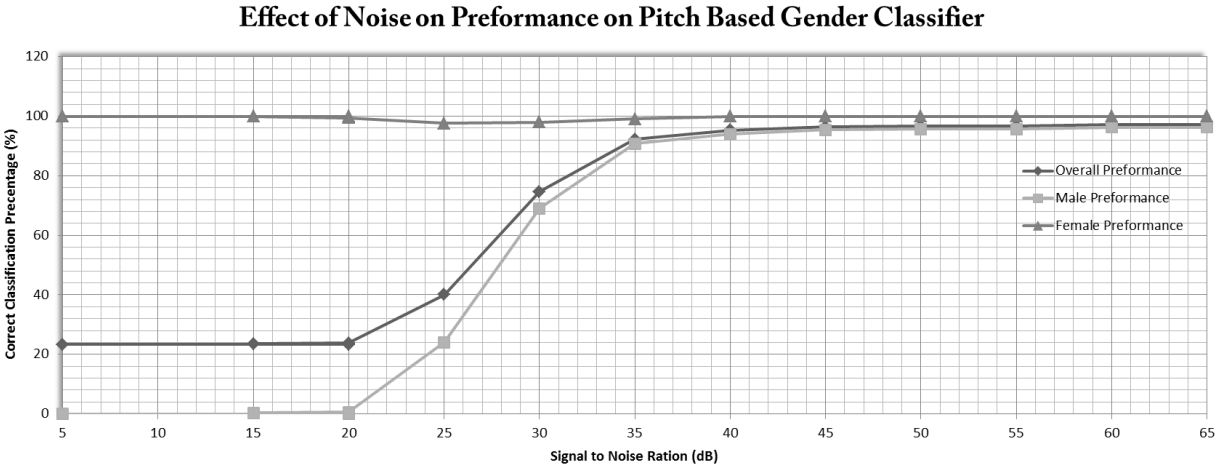


Figure 10: Graph showing the Effect of Noise on Pitch based Gender Classification

Generally a speech signal with SNR above 30 dB is considered a clean signal while a human will barely notice anything better than a SNR of 20 dB [13]. Despite this, classification performance starts to drop at an SNR of 55 dB. Performance also starts to drop dramatically once SNR is below 35 dB.

Below 30 dB the YIN algorithm is almost completely unable to accurately estimate pitch. A point to note is that the correct classification of female utterance stays at 100 % despite increasing SNR, while classification of male utterances drops to 0 %. This is due to the increased noise causing much larger pitch estimates which results in all utterances being classified as female.

Two approaches to dealing with the influence of noise would be to pre-process the speech signal to reduce the effects of the noise and to allow for better pitch estimations. Secondly the training data should also contain the same level of noise so as to give an accurate representation of the utterances we're testing.

3.2. Classification Using MFCCs

Despite suggestions (shown in section 2.3) that MFCCs would be a poor feature for gender classification an MFCC gender classifier was designed and tested. As with the pitch based gender classifier it is difficult to know how many components should be used for each GMM. Additionally we must choose how many MFCCs to use per windowed section. Documentation on the subject of the number of MFCCs disagrees, with paper [9] suggesting that increasing MFCCs continues to increase performance (with 12 being the optimum when considering computational performance) while paper [8] suggests that at a certain point performance starts to drop off with increased number of coefficients (again suggesting that 12 is the optimum). Figure 11 shows the effects of number of MFCCs and number of mixture components on the performance of the MFCC based gender classifier. Tables 8, 9 and 10 show the details of classification performance for the different numbers of mixture components and MFCCs.

The results shown in tables 8, 9 and 10 and in figure 11 are the average of three simulations conducted using the second set of test data (150 male and 150 female utterances). In the majority of cases there was no change in classification rate between simulations except when the GMM failed to converge within 1,000 iterations. There are several important points to note here. Firstly (and most importantly) the gender classifier using MFCCs as features has been shown to have a very high classification rate (above 90%). This contradicts the findings shown in section 2.3 that suggests MFCCs for male and female speech are inseparable. This is possibly due to the temporal effects of the MFCCs. MFCCs are used to model the shape of the vocal tract of the of the human speech system. The vocal tract shape changes dramatically depending on the phoneme being spoken. As a result, plotting the MFCCs for an entire utterance results in a wide distribution which is why the male and female MFCCs appear to overlap. If we were to plot the MFCCs for a male and female uttering the same phoneme the MFCCs may prove to be more separable. This is a distinctive feature of MFCCs and is the reason they are commonly used along with HMMs in systems for speech and speaker recognition (i.e. systems that are modelled based on temporal effects). The result is that by sectioning the utterance and classifying each section separately we get a more reliable classification. The overall classification of the utterance is then taken as the class which has the most sections of the utterance. Comparatively taking the average PDF value of all the sections of an utterance gives a classification rate of around 65%.

	Mixture 1	Mixture 2	Mixture 3	Mixture 4	Mixture 5	Mixture 6	Mixture 7	Mixture 8	Mixture 9	Mixture 10
Overall Percentage	73.00 %	83.33 %	88.67 %	90.33 %	89.33 %	92.00 %	86.67 %	94.67 %	92.67 %	95.33 %
Incorrect Male	9	15	6	3	1	3	1	2	1	3
Incorrect Female	72	35	28	26	31	21	39	14	21	11
Male Percentage	94.00 %	90.00 %	96.00 %	98.00 %	99.33 %	98.00 %	99.33 %	98.67 %	99.33 %	98.00 %
Female Percentage	52.00 %	76.67 %	81.33 %	82.67 %	79.33 %	86.00 %	74.00 %	90.67 %	86.00 %	92.67 %

Table: 8 Classification performance of male and female utterances for different number of GMM components when using 6 MFCC

	Mixture 1	Mixture 2	Mixture 3	Mixture 4	Mixture 5	Mixture 6	Mixture 7	Mixture 8	Mixture 9	Mixture 10
Overall Percentage	89.00 %	94.00 %	96.00 %	96.33 %	96.00 %	97.33 %	97.67 %	98.00 %	98.67 %	99.00 %
Incorrect Male	9	8	4	2	5	1	2	1	1	0
Incorrect Female	24	10	8	9	7	7	5	5	3	3
Male Percentage	94.00 %	94.67 %	97.33 %	98.67 %	96.67 %	99.33 %	98.67 %	99.33 %	99.33 %	100.00 %
Female Percentage	84.00 %	93.33 %	94.67 %	94.00 %	95.33 %	95.33 %	96.67 %	96.67 %	98.00 %	98.00 %

Table: 9 Classification performance of male and female utterances for different number of GMM components when using 12 MFCC

	Mixture 1	Mixture 2	Mixture 3	Mixture 4	Mixture 5	Mixture 6	Mixture 7	Mixture 8	Mixture 9	Mixture 10
Overall Percentage	96.67 %	98.67 %	99.33 %	98.67 %	98.33 %	99.33 %	98.67 %	98.67 %	99.33 %	99.67 %
Incorrect Male	1	0	0	0	0	0	0	0	0	0
Incorrect Female	9	4	2	4	5	2	4	4	2	1
Male Percentage	99.33 %	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %
Female Percentage	94.00 %	97.33 %	98.67 %	97.33 %	96.67 %	98.67 %	97.33 %	97.33 %	98.67 %	99.33 %

Table: 10 Classification performance of male and female utterances for different number of GMM components when using 24 MFCC

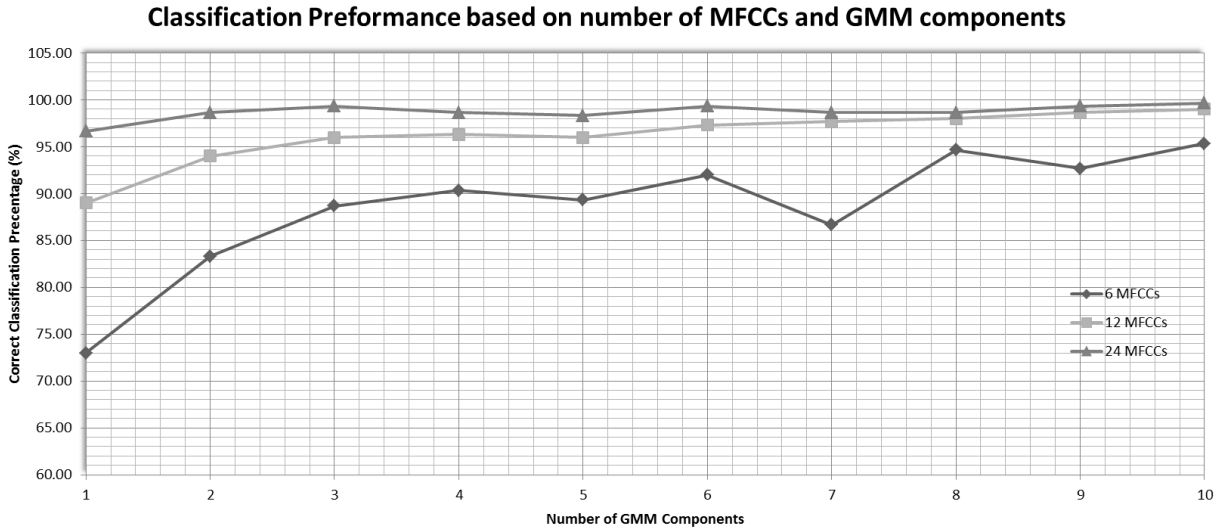


Figure 11: Graph showing the effect the number of GMM components has on classification when using 6, 12 and 24 MFCCs

The second point to note from figure 11 is that classification accuracy increases with increased number of MFCCs. This shouldn't be too surprising since the MFCCs are uncorrelated, so increasing the dimensionality of the GMM may help to separate the male and female classes. This can be somewhat seen by comparing figures 5 and 6 where we can see that there is less overlap in the case of bivariate data than univariate data.

The final point to note is the effect that increasing the number of mixtures has on the classification performance. In the case of 6 and 12 MFCCs, the performance seems to level out above 3 mixture components. However for 6 MFCCs the performance seems to continue increasing until around 8 mixture components. This is interesting since usually it would be easier to fit a GMM to a feature set with lower dimensionality. Meaning we should achieve the same results with fewer components. Figure 11 suggests that this isn't the case

Unfortunately MFCC features also suffer from the same issue as pitch features, which is that they are highly sensitive to noise. Again to demonstrate this, white Gaussian noise was added to each of the test utterances before calculating the MFCC features. Figure 12 shows the effect of noise on classification performance with noise signals ranging from 5 dB to 65 dB.

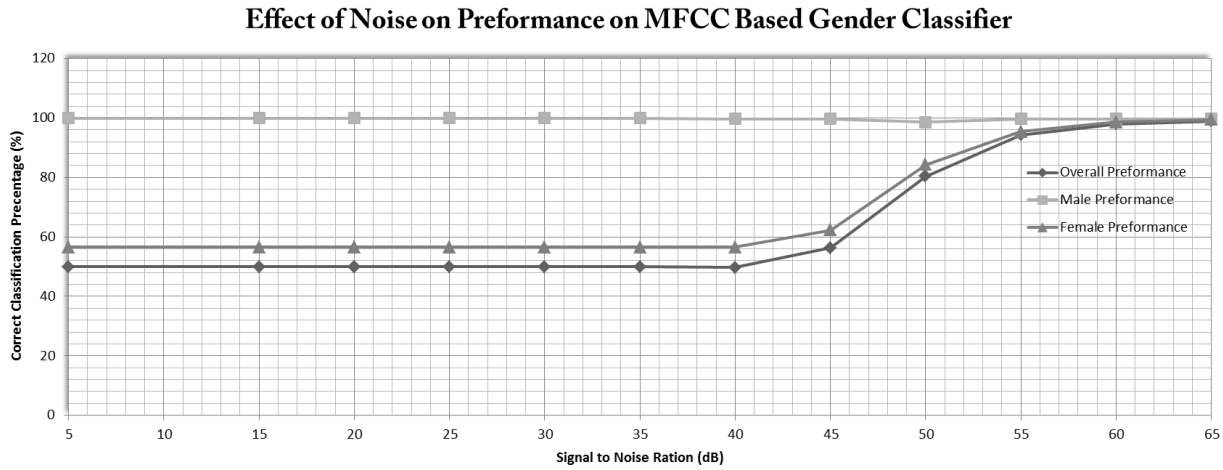


Figure 12: Graph showing the Effect of Noise on MFCC based Gender Classification (12 MFCC and 6 mixture components)

As mentioned above an SNR ratio of 30 dB is considered a clean speech signal. Additionally there is almost no detectable drop in quality to a human listener for speech signals with SNR above 20 dB. Despite this we can see from figure 12 that classification of the MFCC based gender classifier begins to drop for an SNR of less than 55 dB. The interesting point to note here is that the effect of added noise on the MFCC calculations is causing every utterance to be classified as male (as opposed to the effect of noise on pitch based classification which causes classification of every utterance as female).

The issues MFCCs have with signal noise is well documented [8]. A couple of possible solutions to dealing with noisy signals would be to pre-process the speech signal to reduce the effects of the noise and to allow for clean MFCC calculations. Additionally the training data should also contain the same level of noise so as to give an accurate representation of the utterances we're testing.

3.3. Combined Pitch and MFCC Based Classification

Paper [5] discusses the use of combining Pitch and MFCC features to create a stronger gender classifier. Since both pitch and MFCC feature based classification has already been implemented (trained Respective GMMs) it seems logical to combine the two to test for improved performance. The algorithm was implemented as follows:

1. Since pitch based classification is faster than MFCC classification we use this for the majority of utterances. Based on the average pitch value from the test utterance, the PDF value for both male and female GMMs (pitch) is calculated.
2. If there is a large difference between these two values (i.e. greater than one fifth the maximum height of the male GMM) then classification is done using just the pitch. Otherwise classification is done using MFCCs. The idea is that the utterances that are misclassified by a pitch based classifier may be correctly classified by an MFCC classifier.
3. Running the test data through the combined pitch and MFCC classifier using GMMs calculated from 12 MFCCs and 6 mixtures gives a classification rate of 99.67 %. This is an improvement on both the pitch based classifier (97.4% for 6 mixtures) and the MFCC based classifier (97.33% for 6 mixtures). Although this improvement is slight there is a second improvement which is slightly faster runtime compared to MFCC based classification.

3.4. Possible Improvements

As mentioned in section 2.5 the primary issue with using pitch as a feature for gender classification is the difficulty of obtaining an accurate pitch estimate from a sample utterance. This is shown in figure 10 to be particularly prominent in the presence of noise. A possible work around for this would be to pre-process the speech to reduce the effects of noise. The main issue here however is not with pitch as a feature but with our ability to determine pitch.

Similarly, MFCCs are also very sensitive to noise and as a result have poor performance when noisy signals are used. Since MFCCs have shown to produce strong classification results it would be worth trying power normalised Cepstral Coefficients (PNCC) which preform a similar function to MFCCs but are more robust to noise. Additionally the noise could be dealt with by training the respective GMMs with utterances also corrupted by the same level of noise. This is unlikely to produce much of an improvement in the case of pitch based classification. However, as the industry advances, more robust pitch based algorithms are likely to be found suggesting that pitch based classification could become a highly robust method of determining gender from speech.

However greater improvements might be seen through the use of different features or a different clustering structure such as an SVM as suggested in [7]. As mentioned in section 2.5 possible alternative features could include models of speaker rhythm, tempo, timbre and stress. Additionally some high level paralinguistic features have been shown to be able to distinguish male and female speakers [17] however, calculation of these features would require accurate word recognition and understanding.

4. Conclusion

In this report we have shown that male and female utterances can be almost entirely separated based purely on fundamental frequency. A gender classification system was designed using GMMs and fundamental frequency as a classification feature. The classification system showed performance of above 95 % for both male and female utterances when clean signals were used. The classification rate greatly dropped off however in the presence of even a small amount of noise. Similarly a gender classification system was designed using MFCC coefficients. Despite suggestions that MFCCs may prove inseparable when applied to the problem of gender identification the classification rate was shown to be very high (around 97 %). However MFCCs also show a large drop in performance when noisy signals are used. Finally a classifier was designed using both pitch and MFCCs which showed a slight improvement in both classification performance and computation time compared to a simple MFCC classifier.

5. References

[1] *Kim, Hye-Jin, Kyungsuk Bae, and Ho-Sub Yoon. "Age and gender classification for a home-robot service." Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on. IEEE, 2007.*

[2] *Tzanetakis, George, and Perry Cook. "Musical genre classification of audio signals." Speech and Audio Processing, IEEE transactions on 10.5 (2002): 293-302.*

[3] Harb, Hadi, and Liming Chen. "Voice-based gender identification in multimedia applications." Journal of intelligent information systems 24.2-3 (2005): 179-198.

[4] Hu, Yakun, Dapeng Wu, and Antonio Nucci. "Pitch-based gender identification with two-stage classification." Security and Communication Networks 5.2 (2012): 211-225.

[5] Ting, Huang, Yang Yingchun, and Wu Zhaohui. "Combining MFCC and pitch to enhance the performance of the gender recognition." Signal Processing, 2006 8th International Conference on. Vol. 1. IEEE, 2006.

[6] Reynolds, Douglas. "Gaussian mixture models." Encyclopedia of Biometrics. Springer US, 2009. 659-663.

[7] Fokoue, Ernest, and Zichen Ma. "Speaker Gender Recognition via MFCCs and SVMs." (2013).

[8] Practicle Cryptography. "Mel Frequency Cepstral Coefficient (MFCC) tutorial". Accessed at <http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/> on 30/12/2015.

[9] Yücesoy, Ergün, and Vasif V. Nabiyev. "Gender identification of a speaker using MFCC and GMM." Electrical and Electronics Engineering (ELECO), 2013 8th International Conference on. IEEE, 2013.

[10] YIN MatLab Code was used in the testing of this algorithm the code used was written by Alain De Cheveigne and was obtained on 1/11/2015 at http://audition.ens.fr/adc/

[11] Code used for calculating the MFCCs was taken from the Voice box library obtained on 28/10/2015 at http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[12] Alternative MFCC code written by Kamil Wojcicki on 10/12/2015 at http://www.mathworks.com/matlabcentral/fileexchange/32849-htk-mfcc-matlab/content/mfcc/mfcc.m

[13] Dave Gelbart and George Doddington. "How is the SNR of a speech example defined"(2000) access at <http://www1.icsi.berkeley.edu/Speech/faq/speechSNR.html> on 2/1/2016

[14] Glosup, J. G., and M. C. Axelrod. Use of the AIC with the EM algorithm: A demonstration of a probability model selection technique. No. UCRL-JC--118762; CONF-9408107--6. Lawrence Livermore National Lab., CA (United States), 1994.

[15] Santos, Timothy Israel, and Rowena Cristina Guevara. "Classification of Filipino Speech Rhythm Using Computational and Perceptual Approach." PACLIC. 2011.

[16] Aucouturier, Jean-Julien, François Pachet, and Mark Sandler. "" The way it Sounds": timbre models for analysis and retrieval of music signals." Multimedia, IEEE Transactions on 7.6 (2005): 1028-1035.

[17] Attenborough, Frederick. "Words, contexts, politics." Gender and Language 8.2 (2014): 137-146.

[18] Jebara,Tony, Wenwei Wang. "Voice-Based Gender Classification Using Support Vector Machine". (2006). Columbia University Presentation

[19] Kim, Chanwoo, and Richard M. Stern. "Power-normalized cepstral coefficients (PNCC) for robust speech recognition." Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012.