Automatic Gender Classification Using the Mel Frequency Cepstrum of Neutral and Whispered Speech: a Comparative Study

G. Nisha Meenakshi  
Electrical Engineering  
Indian Institute of Science (IISc)  
Bangalore 560 012, India  
Email: gnisha@ee.iisc.ernet.in  

Prasanta Kumar Ghosh  
Electrical Engineering  
Indian Institute of Science (IISc)  
Bangalore 560 012, India  
Email: prasantg@ee.iisc.ernet.in

Abstract—A whispered speech resembles an unvoiced speech due to the lack of vocal fold vibration unlike the neutral speech. Since information about the gender of a speaker typically lies in the pitch resulted from the vocal fold vibration (or source signal), identifying gender from the whispered speech is more challenging compared to that from the neutral speech. In the absence of the pitch, we study the use of the vocal tract filter captured through the spectral envelope for automatic gender classification (AGC) from a whispered speech. The spectral envelope is represented by the Mel frequency cepstral coefficients (MFCCs). We also compare the AGC performance from the neutral speech using only MFCCs with that from the whispered speech. AGC experiment using a set of 33 sentences spoken in neutral and whispered mode by 16 female and 20 male speakers reveals that the AGC accuracy using the neutral speech is, on average, higher (4.5% absolute) than that using the whispered speech when only the spectral shape information is used. This is true even when we use a subset of MFCCs obtained by a forward cepstral coefficient selection algorithm. However, the AGC accuracy obtained using the MFCC of the neutral speech is found to be 2.83% (absolute) lower compared to that using pitch. These findings not only suggest that there is gender specific information in the spectral shape but also indicate that the spectral shape carries less gender specific information when a speaker whispers as opposed to speaking normally.

I. INTRODUCTION

Whispering is a natural form of communication that emerges in private as well as pathological situations. The whispered speech lacks the pitch information due to absence of vibration of the vocal folds [1]. Several algorithms have been developed for reconstructing neutral speech from the whispered speech which involve an essential pitch estimation and a pitch insertion procedure [2], [3]. Automatic gender classification (AGC) from whispered speech could help in making the appropriate choice of pitch for insertion in such reconstruction algorithms.

In the late 1960s, Schwartz [4] and Ingemann [5] investigated speaker’s gender identification from voiceless fricatives, via listening tests. It was reported that the difference in shift in the formant frequencies for the female and male speech is an important cue to gender identification [4]. It was argued that the presence of strong formant-like spectral peaks in the spectrum of certain voiceless fricatives enhances better gender identification performance [5]. Similar studies on the whispered vowels were carried out by Schwartz et al. [6] and Lass et al. [7]. From studies on neutral speech it was found that the laryngeal fundamental is more indicative of the speaker’s gender compared to the vocal tract information i.e., the formants [7]. It is to be noted that these investigations are based on listening tests.

Automatic gender classification (AGC) from neutral speech is a well-explored area of research. Detailed studies on the feature vectors and distance measures for AGC from neutral speech have been thoroughly done [8], [9]. Gender classification from children’s speech and gender identification in the context of multimedia indexing has also been addressed [10], [11]. Pitch, the primary indicator of speakers’ gender, has been used for language independent gender classification [12]. It was reported that combining the relative spectral perceptual linear predictive coefficients (RASTA-PLP) with parameters of pitch, yields good AGC performance even in the presence of noise [13]. The vocal tract information captured by the Mel frequency cepstral coefficients (MFCCs) has also been used to distinguish male and female speech [14]. A fusion scheme to combine both the pitch based features and the MFCC based features was devised [15] to improve the AGC performance. Using this fusion scheme, the performance of the binary Support Vector Machine (SVM) classifier was found to be better than the Gaussian mixture model (GMM) framework for an AGC from the neutral speech.

To the best of our knowledge, automated algorithms for gender identification from whispered speech are not explored well in the literature. Since the whispering is noise excited, it lacks the periodic excitation and hence the pitch is missing from a whispered speech [1]. Therefore, in this work, we would like to explore the potential of the vocal tract information from the whispered speech for AGC. The spectrum of the whispered speech is known to have formant shifts at lower frequencies [16]. The study of the vowel spaces of whispered speech reveals that there is a difference in the shape of the acoustic vowel diagram of male whisper and female whisper [17]. Therefore, it is seen that the lower formant shifts affect the spectrum of male and female speech differently. Motivated by this observation, in this paper, we use the vocal tract information present in the whispered speech to automatically...
identify the gender of a speaker from the whispered speech. We use the MFCC vector to capture the envelope of the whispered speech spectrum. In this paper, we analyze the AGC performance using the 13 dimensional static MFCC vector and the 39 dimensional MFCC vector obtained by appending the velocity and the acceleration coefficients to the 13 dimensional vector. In addition to this, we also use a forward cepstral feature selection algorithm, to choose the optimal number of coefficients necessary to maximally discriminate male whisper from female whisper. Since it is well-known that the spectral envelope differs in whispering compared to neutral speech [16], we also perform a comparative study by conducting AGC experiment when the same set of sentences are spoken in a neutral mode. The goal of this study is to quantify how much gender specific information is encoded in spectral shape and how that changes from neutral to whispered speech. We further investigate the AGC performance from the whispered speech when the neutral speech is used to train the classifier as opposed to the whispered speech. This is also repeated for the neutral speech. AGC with train-test mismatched condition is performed to understand the robustness of the spectral shape based representation using neutral and whispered speech. Finally we perform pitch based AGC using neutral speech to examine the difference in AGC performance using pitch and spectral envelope in both the neutral and the whispered speech. From this comparative study, we find that pitch yields higher AGC performance compared to the spectral shape based features. However, spectral shape results in a better AGC accuracy when the neutral speech is used as opposed to the whispered speech. We begin by explaining the dataset used for this study.

II. DATASET

CHAINS (Characterizing Individual Speaker) is a speech corpus [18] comprising 6 speaking conditions, from which whispered speech and neutral (SOLO) speech are used for the present study. The CHAINS corpus consists of 3 different accents of speech from a total of 36 speakers (20 males and 16 females). We consider the 33 sentences spoken by each of the speakers, in whispered and neutral mode. While 24 of these sentences belong to the TIMIT corpus, 9 of them are taken from the CSLU Speaker Identification Corpus. The data is downsampled to 16000Hz for this work. Thus, the total duration of the whispered speech and neutral speech considered for the current study are ∼3477.495 seconds (57.9 minutes) and ∼3317.170 seconds (55.28 minutes) respectively. As an example, Figure 2 shows the whispered speech signal and the corresponding spectrogram of a male subject (Figure 2(a) and (b)) and those of a female subject (Figure 2(c) and (d)) for the sentence ‘She had your dark suit in greasy wash water all year’.

III. AUTOMATIC GENDER CLASSIFICATION

The Figure 1 shows the block diagram illustrating the steps involved in the AGC using MFCC. For each speaking condition, namely, neutral and whisper, the database is divided into the training set, the development set and the test set. At first, K-dimensional MFCCs of the utterances in the training and development sets are computed with 20ms frame duration and 10ms frame shift. MFCCs of all frames are stacked next to each other to form two MFCC matrices for the training and development sets denoted by $M_t$ and $M_d$ respectively. These MFCC matrices are supplied to the forward cepstral feature selection algorithm to obtain the ranked list of MFCCs and to choose the optimal number of MFCCs for each frame. In this paper, we analyze the AGC performance from the whispered speech spectrum. In this paper, we analyze the AGC performance from the whispered speech spectrum. In this paper, we analyze the AGC performance from the whispered speech spectrum.
A subset of MFCCs is used from the ranked list of selected coefficients to create the training features and the test features. A frame-level SVM classifier is trained using the training features. The trained SVM classifier is used to classify each of the test features. A majority voting strategy is then employed over labels obtained using SVM classifier in all the frames of a test utterance to finally determine if the utterance is spoken by a male or a female speaker. We now describe the procedure to obtain the ranked list of MFCCs using forward cepstral feature selection method.

### A. Forward Cepstral Feature Selection

The motivation behind obtaining $C^*$ is to find a set of MFCCs that are rich in the speaker’s gender information. Thus, the goal is to achieve a good gender classification performance from a subset of MFCCs. Let the signals in the training and development sets be denoted by $x_t[n]$ and $x_d[n]$ respectively. To find $C^*$ that would give the maximum accuracy of gender classification on the development set, we follow the forward cepstral feature selection algorithm given in Algorithm 1. In every stage of this iterative algorithm, the chosen set of cepstral coefficients is used to create the training set, which is used to train the SVM classifier, which is finally used to perform a frame-level classification on the development set. In the $j^{th}$ iteration of the feature selection, the AGC accuracy $A_j$ on the development set is found by a majority rule on the labels at each frame given by the SVM classifier. Following this process, in every iteration, the cepstral coefficient that yields the maximum AGC accuracy on the development set is chosen and added to the ranked list of MFCCs.

**Algorithm 1: Forward Cepstral Feature Selection**

1: Inputs: $K, x_t[n], x_d[n]$
2: Initialization: $\tilde{C} = [], \tilde{A} = [], \tilde{C^G} = [1, 2, ..., K]$
3: for $i = 1$ to $K$ do
4:     $A^* \leftarrow \tilde{A}$
5:     for $j \in \tilde{C}^G$ do
6:         $C = [\tilde{C}, j]$
7:         $M_t \leftarrow x_t[n]$
8:         $M_d \leftarrow x_d[n]$
9:         $A_j \leftarrow M_t, M_d$ (using SVM classification)
10:     end for
11:     $j^* = \arg\max_j A_j$
12:     $\tilde{C}_i \leftarrow j^*$ ($\tilde{C}_i$ denotes the $i^{th}$ element of the vector $\tilde{C}$)
13:     $\tilde{C^G} \leftarrow \tilde{C}^G \setminus \{j^*\}$
14:     $\tilde{A}_i \leftarrow A_{j^*}$
15: end for
16: $i^* \leftarrow \arg\max_i \tilde{A}_i$
17: $C^* = \tilde{C}_i, i = [1, ..., i^*]$
18: $A^* \leftarrow \tilde{A}_{i^*}$
19: return $C^*, A^*, \tilde{A}, \tilde{C}$

### IV. EXPERIMENTS AND RESULTS

#### A. Experimental Setup

The data from 36 speakers of the CHAINS corpus, is grouped to create the training set, development set and the test set in each fold of a five-fold cross validation setup, in each speaking condition (neutral and whisper). Each fold contains both male and female speakers in proportion with the total number of male and female speakers in the CHAINS corpus. Each fold is used as the test set while the training and development sets are formed using remaining four folds. This is repeated five times in a round-robin fashion. In the neutral condition, the average durations of the training set, the development set and the test set are $\sim 1988.3$ seconds (33.13 minutes), $\sim 602.7$ seconds (11.04 minutes) and $\sim 602.7$ seconds (11.04 minutes) respectively. In the whispered condition, the average durations of the training set, the development set and the test set are $\sim 2086.5$ seconds (34.77 minutes), $\sim 695.49$ seconds (11.59 minutes) and $\sim 695.49$ seconds (11.59 minutes) respectively. The training and the development sets in each fold are used to obtain the $C^*$ using the forward cepstral feature selection algorithm (Algorithm 1). This is then used to perform the gender classification separately on each of the utterances of the test set.

In this study, we first analyze the AGC performance using all the $K = 13$ and $K = 39$ MFCCs. It is to be noted that the $13^{th}$ coefficient of the MFCC vector, represents the dc component, i.e., the signal energy in the respective frame. Next, we perform the gender classification based on the top few selected MFCCs, to examine if comparable classification performance can be achieved using a subset of $K$ coefficients. We also study the effects of training the classifier in one speaking condition and testing in the other, to understand the difference in the amount of gender information present in the two speaking conditions. Finally, we compare the performance of the pitch based AGC with the MFCC based one. The results of these experiments are described in the following subsection.

**Fig. 3.** AGC accuracy on the test set using 13 and 39-dim MFCCs for both whispered and neutral speech for each of five folds.

#### B. Results and Discussion

1) **Gender classification using all MFCCs:** Figure 3 shows the AGC accuracy from the neutral and whispered speech using MFCCs with dynamic features (39-dimensional) as well as without dynamic features (13-dimensional). It is clear that there is no significant difference between the AGC accuracies using 13-dim and 39-dim MFCCs for both neutral and whispered speech. This suggests that there is no additional gender specific information in the dynamic features as opposed to the static features. Hence, for the rest of the experiments we continue using 13-dimensional MFCCs. From Figure 3, it is also
clear that the AGC accuracy using the neutral speech is higher than that using the whispered speech in four folds among five folds considered. This in turn suggests that the spectral envelope in the neutral speech more accurately encodes gender information as opposed to that of the whispered speech. On average, the AGC accuracy using whispered speech is 4.5% lower than that using the neutral speech. This could be because of the fact that formants of many voiced sounds change during whispering leading to different spectral envelopes in neutral and whispered mode of speaking.

2) MFCCs selection for AGC: Figure 4 shows the ranked 13 cepstral coefficients in an order of the coefficient selection using Algorithm 1 for both neutral and whispered speech. The highest AGC accuracy (indicated by a red square) on the development set is obtained using a subset of the MFCCs. This suggests that the key information about gender is encoded in a fewer cepstral coefficients. The highest AGC accuracy is also reported (in red) in each figure corresponding to every fold. It is clear that the AGC accuracy using neutral speech is more than that using whispered speech in every fold. It is also interesting to note that the number of cepstral coefficients corresponding to the highest AGC accuracy in each fold is lower in the case of neutral speech compared to the whispered speech. This suggests that the gender information is more compactly encoded in a fewer MFCCs of the neutral speech as opposed to those of the whispered speech. It is seen that the 1st and 11th coefficients consistently appear in the top five selected MFCCs for all folds of neutral speech. Similarly the 1st coefficient is consistently selected in the top five MFCCs for all folds of whispered speech. This indicates that the 1st MFCC, which corresponds to the logarithm of the ratio of the weighted sum of energies in low frequencies to that of the high frequencies, carries cue to gender information of a speaker for both neutral and whispered speech. In case of the neutral speech, 4th and 9th MFCCs are in top five selected coefficients in three among five folds. Similarly, in case of the whispered speech, 5th coefficient and 3rd coefficient appear in the top five selected coefficients in respectively four and three among five folds. Thus, except the 1st MFCC, the top few selected MFCCs are different for whispered and neutral speech.

3) Gender classification using selected MFCCs: We perform AGC experiment on the test set using top few selected MFCCs obtained from the development set. This is done to examine the robustness of the selected features on an unseen test set. We consider three different subsets of features along with the complete 13-dimensional MFCCs for this purpose. In particular, we select to 25%, 50%, and 75% top selected features from the ranked cepstral coefficients obtained in the previous Section IV-B2. The AGC accuracy on the test set of each fold is listed in Table I for both neutral and whispered speech. It is clear that the AGC accuracy on the test set increases with increase in the percentage of the selected coefficients. However, the AGC accuracy does not consistently increase from 75% selected features to the entire (100%) MFCCs for all folds. This is true for both neutral and whispered case. In fact, the average improvement is 0.46% and 1.17% for neutral and whispered speech respectively. Thus, 75% of the selected features can be used in place of the entire MFCCs without any significant loss in the AGC accuracy. It is also clear from Table I that for each of the coefficient subsets, the average AGC accuracy is always higher using the neutral speech compared to the whispered speech. This is consistent with the accuracy obtained using the entire set of MFCCs.

### Table I. Accuracy (%) of Gender Classification Using a Percentage of Selected MFCCs for All Five Folds of Neutral and Whispered Speech

<table>
<thead>
<tr>
<th>% Feature</th>
<th>Speaking condition</th>
<th>Fold 1</th>
<th>Fold 2</th>
<th>Fold 3</th>
<th>Fold 4</th>
<th>Fold 5</th>
<th>Avg (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>Neutral</td>
<td>60.98</td>
<td>62.34</td>
<td>96.10</td>
<td>88.74</td>
<td>65.37</td>
<td>74.71 (14.46)</td>
</tr>
<tr>
<td></td>
<td>Whisper</td>
<td>52.65</td>
<td>61.90</td>
<td>75.75</td>
<td>83.54</td>
<td>58.87</td>
<td>66.54 (12.72)</td>
</tr>
<tr>
<td>50</td>
<td>Neutral</td>
<td>85.22</td>
<td>91.50</td>
<td>98.83</td>
<td>95.70</td>
<td>78.78</td>
<td>90.81 (8.58)</td>
</tr>
<tr>
<td></td>
<td>Whisper</td>
<td>87.50</td>
<td>94.45</td>
<td>92.20</td>
<td>93.98</td>
<td>68.84</td>
<td>81.39 (9.57)</td>
</tr>
<tr>
<td>75</td>
<td>Neutral</td>
<td>89.01</td>
<td>94.40</td>
<td>99.56</td>
<td>96.96</td>
<td>83.98</td>
<td>93.48 (6.60)</td>
</tr>
<tr>
<td></td>
<td>Whisper</td>
<td>95.45</td>
<td>96.14</td>
<td>97.40</td>
<td>89.17</td>
<td>73.16</td>
<td>89.26 (9.69)</td>
</tr>
<tr>
<td>100</td>
<td>Neutral</td>
<td>88.20</td>
<td>93.27</td>
<td>99.57</td>
<td>100.00</td>
<td>81.82</td>
<td>93.84 (8.46)</td>
</tr>
<tr>
<td></td>
<td>Whisper</td>
<td>96.96</td>
<td>96.14</td>
<td>95.67</td>
<td>91.74</td>
<td>77.00</td>
<td>89.43 (8.11)</td>
</tr>
</tbody>
</table>
indicating that at any given subset of MFCCs, the neutral speech carries more gender specific information compared to that of the whispered speech.

4) AGC with mismatched train-test sets: Spectral shape of voiced sounds in neutral speech is different from that in the whispered speech [16]. From the AGC experiments with neutral and whispered speech separately, we observe that the spectral shapes of the neutral speech sounds carry more gender specific information compared to those of the whispered speech. However, it is not clear if an AGC system, trained with the neutral speech, would do equally well when the whispered speech is used as the test set and vice-versa. In order to examine this, we conduct two additional experiments: 1) neutral speech for training and whispered speech as the test set, 2) whispered speech for training and neutral speech as the test set. The AGC accuracies under these two setups are shown in Table II and III. MFCCs of 13 as well as 39 dimensions are used for this purpose. It is clear from Tables II and III that the AGC accuracy on the whispered test set with neutral speech for training is higher than that on the neutral test set with whispered speech for training. This is true for both 13 and 39 dimensional MFCC features. However, the performance between the 13 and 39-dim MFCCs cases are similar in each of the two setups. These results indicate that although the spectral shapes in the whispered speech may be different from those in the neutral speech, the gender classifier trained using neutral speech is more robust compared to that trained using whispered speech.

5) Comparison of pitch and MFCC based AGC: From Table I, we see that the best average AGC accuracy from both the neutral as well as the whispered speech is obtained using the 13 MFCCs and the accuracy is higher for the neutral (93.84%) compared to the whispered speech (89.43%). Since the gender information is reliably represented by the pitch of a person, we also run a pitch based AGC experiment using voiced frames of the neutral speech in an identical five-fold cross validation setup. The average AGC accuracy using pitch turns out to be 96.67% (±3.04%), i.e., 2.83% (absolute) higher than the spectral shape based AGC accuracy. This result suggests that the poor AGC accuracy using whispered speech happens not only because of the missing pitch information but also due to the fact that the spectral shapes (vocal tract filter) in whispered sounds has poor representation capacity compared to those in the neutral speech for gender discrimination.

V. CONCLUSIONS

We use a MFCC-SVM framework to perform automatic gender classification from whispered speech. We devise a forward cepstral feature selection algorithm that selects a subset of MFCCs that maximally discriminate female whisper from male whisper. We also perform AGC from the neutral speech using MFCC-SVM approach. We find that there no significant additional speaker’s gender information present in the dynamic components in MFCCs compared to the static components in MFCCs. This is true for both neutral as well as whispered speech. Using all 13 static dimensions of the MFCC vector, we observe a higher AGC performance in the neutral speech compared to the whispered speech indicating better cues provided by the spectral shape of neutral speech compared to whispered speech. Using a forward cepstral coefficient selection algorithm we find that a subset of the MFCCs are capable of resulting in an AGC performance similar to that using all MFCCs. Further, it is observed that the 1st MFCC is rich in gender specific information since it is selected as the top coefficient for both neutral and whispered speech. We also find that the AGC performance on neutral speech degrades when trained on the whispered speech. This degradation is more compared to a setup where the neutral speech is used for training and the whispered speech is used for testing. We have also found that the MFCC based AGC performance is lower than that using pitch from neutral speech. However, it will be interesting to examine the robustness of the pitch and MFCC for AGC in the presence of noise. These are parts of our future work.

REFERENCES


