Analysis of F0 and Cepstral Features for Robust Automatic Gender Recognition

Marianna Pronobis[1,2]*, Mathew Magimai.-Doss*[1]

[1] Idiap Research Institute, Martigny, Switzerland
[2] School of Electrical Engineering, Royal Institute of Technology (KTH), Sweden
madry@kth.se, mathew@idiap.ch

Abstract

In this paper, we analyze applicability of F0 and cepstral features, namely LPCCs, MFCCs, PLPs for robust Automatic Gender Recognition (AGR). Through gender recognition studies on BANCA corpus comprising datasets of varying complexity, we show that use of voiced speech frames and modelling of higher spectral detail (i.e. using higher order cepstral coefficients) along with the use of dynamic features improve the robustness of the system towards mismatched training and test conditions. Moreover, our study shows that for matched clean training and test conditions and for multi-condition training, the AGR system is less sensitive to the order of cepstral coefficients and the use of dynamic features gives little-to-no gain. F0 and cepstral features perform equally well under clean conditions, however under noisy conditions cepstral features yield robust system compared to F0-based system.

Index Terms: automatic gender recognition, cepstral coefficients, fundamental frequency, robustness

1. Introduction

Given an audio/speech signal, the goal of *Automatic Gender Recognition* (AGR) is to identify the gender of the speaker. The output of the AGR system can be useful for different applications, such as building gender specific acoustic models for automatic speech recognition, reducing the search space in speaker recognition or surveillance systems, analyzing human-computer interaction, and social interaction and behaviour.

In the literature, different feature representations for AGR were studied, such as fundamental frequency (F0), formants with their respective frequency, amplitude and bandwidth, linear prediction cepstral coefficients (LPCCs), mel frequency cepstral coefficients (MFCCs). Most of the previous studies on AGR mainly analyzed these features for specific phonemes or broad phonetic classes and under clean conditions.

In this paper, we analyze F0 and three different cepstral features: LPCCs, MFCCs and PLPs (perceptual linear prediction coefficients) for AGR under varying conditions. During the analysis, we also address specific questions such as:

1. What is the effect of data selection (selection of particular frames) on the AGR accuracy?
2. What is the effect of the cepstral feature dimension on the AGR accuracy?
3. What is the effect of training data conditions on the AGR accuracy?

We performed the studies on the BANCA database comprising datasets of varying complexity [1]. Our studies show that selection of voiced speech frames helps in yielding a better AGR system, especially for noisy conditions. Moreover, the AGR system employing F0 can attain better performance than cepstral-based AGR system in clean conditions. Increasing the cepstral feature dimension has little effect on the accuracy of the system in clean conditions, but leads to significant improvements in noisy conditions. Finally, training the gender classifier with multi-condition data is mainly helpful in case of the cepstral features for highly noisy/adverse conditions.

2. Motivation

It was observed that females usually have shorter and thinner vocal cords than males. As a result, the F0 of female voices is typically higher than F0 of male voices. This makes F0 an obvious choice for gender recognition [2]. The spectral quality of female and male speech also differs due to the fact that females have in average 0.8 times shorter vocal tracts [3]. As a result, the typical female formant pattern is scaled upward in frequency compared to the male pattern. Moreover, it was shown that female spectra have a steeper slope compared to male spectra. Since, cepstral features extracted from the short-term signal capture the smooth spectral information, they can be used for gender recognition.

One of the first extensive works on AGR aimed at identifying the most appropriate features of speech signal for the task. Comparison of F0 and formant features (frequencies, amplitudes and bandwidths) for ten vowels extracted from the clean-condition speech data of 52 speakers was presented in [2]. It was revealed that first four formant frequencies are superior to corresponding formant amplitudes and bandwidths, and besides, formant frequencies are slightly better than F0. Further analysis of different parametric representations of speech signal (linear prediction, autocorrelation, reflection coefficients and cepstrum) was performed on the same database for vowels, voiced and unvoiced fricatives [3]. It was found that cepstral features yield the best system and the performance improves when increasing linear prediction order from 8 to 20. It was also observed that AGR for vowels and voiced fricatives attain better performance than for unvoiced fricatives. Moreover, the study implied that gender information is time invariant, phoneme independent, and speaker independent for a given gender. In [4], 9 initial MFCCs were evaluated for different groups of phonemes. The study showed that AGR based on vowels, nasal, liquids perform better than AGR based on fricatives, stops, and silence and sound 'H'. It was also found that the static coefficients are superior to the first order dynamic (delta) coefficients, and that the use of both types of coefficients (static+delta) may improve performance.

The aforementioned studies were conducted on high quality, data acquired under clean conditions. However, a very limited number of works considered the performance of F0 in re-

alistic scenarios in case of which two practical problems occur. First, the reliability of F0 estimation can be easily affected by existence of low-frequency noise in recordings or any other degradation of speech quality. Second, the value of F0 changes with the physical and emotional state of a subject. Humans, while speaking spontaneously, often raise their F0 in order to stress some parts of utterance or to make their voices more audible in the presence of high level background noise. Thus, the values of F0 obtained under realistic conditions may highly deviate from those pre-allocated to females and males under clean conditions. Similarly, the cepstral features can be affected in realistic scenarios through e.g. environmental variations and background noise, poor quality microphone or speaker varying intensities. This motivated us to revise AGR studies and analyze F0 and cepstral features with emphasis on the following practical questions:

- The aforementioned studies have shown that AGR accuracy is not the same across all groups of phonemes. This information can be exploited to built a better AGR system that, ideally, will identify gender based on a carefully selected group of phonemes. However, in practice, this approach is complex and requires the use of e.g. a phoneme recognizer before AGR. Our approach is based on the observation that typically voiced segments provide the best performance, as shown in [3, 4]. Thus, can selection of voiced frames lead to a better AGR system compared to the approach employing the entire speech segments?
- In [3], it was shown that increase of linear prediction order and of number of cepstral coefficients leads to improvement in the AGR recognition rate. Does this trend hold also under noisy conditions?
- How much does the quality of data influence the gender information present in F0 and the cepstral features? Is it a better strategy to train the AGR system on only clean-condition data or on multi-condition (clean+noisy) data?

3. Experimental Setup

3.1. Database

The AGR system was evaluated on the BANCA database (English corpus) comprising datasets of varying complexity [1]. Data acquisition was performed using two microphones (poor-quality and good-quality) under three different types of conditions: (a) *Controlled*: good-quality microphone, clean conditions; (b) *Degraded*: poor-quality microphone, stable conditions; (c) *Adverse*: good-quality microphone, background noise, arbitrary conditions. In every conditions, 4 sessions were scheduled during which 2 recordings from 52 subjects (26 females, 26 males) were collected. In order to evaluate the system on the same number of known and unknown subjects, only a half of the subjects (26) was used for training. Then, all subjects (52) were divided into two groups consisting of 16 and 36 subjects which were used for development and testing. In order to evaluate performance of the system under clean conditions, the **0** protocol (matched clean training and test conditions) is established. Then, to determine a strategy of training that will yield the most robust system under noisy conditions, three additional protocols: **A**, **B**, **C** were defined each in two versions: *Deg* and *Adv* for degraded and adverse conditions, respectively. As specified in Table 1, the three protocols differ with respect to the quality of data used for training, development and testing. The idea is to first use the controlled conditions data for training and development, and test the system under noisy conditions (protocol **A**, mismatched training and test conditions).

Protocol	Set ID	Conditions			BANCA Session		
		TRAIN	DEV	TEST	TRAIN	DEV	TEST
0	Con_0	**Con.**	**Con.**	**Con.**	1,2	3,4	3,4
A	Deg_A	**Con.**	**Con.**	Deg.	1,2	3,4	7,8
	Adv_A	**Con.**	**Con.**	Adv.	1,2	3,4	11,12
B	Deg_B	**Con.**	Deg.	Deg.	1,2	7,8	7,8
	Adv_B	**Con.**	Adv.	Adv.	1,2	11,12	11,12
C	Deg_C	**Con.**+Deg.	Deg.	Deg.	1,2,5,6	7,8	7,8
	Adv_C	**Con.**+Adv.	Adv.	Adv.	1,2,9,10	11,12	11,12

Protocol	Item	TRAIN	DEV	TEST
0,**A**,**B**,**C**	Subjects Σ(F,M)	26(13,13)	16(8,8)	36(18,18)
	Data per File	1.5s	1.3s	1.3s
0,**A**,**B**	# Files	104	64	144
C	# Files	104+64	64	144

Table 1: Experimental setup for different protocols. Abbreviations and symbols: 'Con.'=Controlled, 'Deg.'=Degraded, 'Adv.'=Adverse, 'Σ'=Total, 'F'=Females, 'M'=Males.

Next, the noisy data are used for development (protocol **B**) and, finally, are added to the training set (protocol **C**, multi-condition training). Finally, we report the results for an utterance based on 1.3s of speech segment or voiced speech segment. The decision about gender for a single utterance was obtained by summing the frame values of the a posteriori probabilities for each class over the whole segment, and then choosing the class with the higher score.

3.2. Analysis of Audio Data

The audio signal was sampled at 16kHz and analyzed in frames of 25ms using a frame shift interval of 10ms. For each utterance, the speech/non-speech segmentation is obtained by first training a GMM with two mixtures. The mixture with largest energy coefficient is labelled as speech and the other as silence, and then followed by the classification of the frames. The RAPT algorithm was used to obtain both F0 estimates and voicing information [5]. The three cepstral features, namely LPCCs, MFCCs and PLPs were extracted using the HTK toolkit, and we analyzed their performance with respect to the number (9, 13, 19) and type (static vs. static+delta) of cepstral coefficients included to the feature vector.

3.3. Classification

We employed the SVMs implemented in the LIBSVM library to perform gender classification [6]. The RBF kernel was used in case of multi-dimensional feature vectors and the linear kernel in case of one-dimensional feature vectors. The parameters of the SVMs (error penalty C) and the RBF kernel (variance γ) were estimated on the development set. The a posteriori class probability, estimated using the Platt's method [7], was chosen a measure of confidence with which a sample was assigned to a particular class.

4. Results and Discussion

In this paper, we present three sets of experiments. First, we compare two different data selection approaches: (a) where speech part of the signal (both voiced and unvoiced frames) is used as a source of information about gender, or (b) where only voiced frames are used instead (Section 4.1). Second, we study the effectiveness of F0 for the AGR problem under varying conditions (Section 4.2). Finally, we study three different types of cepstral features: LPCCs, MFCCs and PLPs (Section 4.3).

4.1. Frame Selection

We report the performances for two systems that differ in the frame selection method in Table 2. In the first system, a

Data Type	Feature	Accuracy [%]		
		Con_0	Deg_A	Adv_A
Speech	F0*	99.3	95.8	91.7
	$LPCC18_\Delta$	94.4	89.6	79.2
	$MFCC19_\Delta$	97.9	91.7	80.6
	$PLP19_\Delta$	95.8	86.1	81.3
Voiced	F0	100.0	95.8	93.1
	$LPCC18_\Delta$	97.2	98.6	96.5
	$MFCC19_\Delta$	97.9	97.9	93.1
	$PLP19_\Delta$	98.6	98.6	97.2

Table 2: Performance of the AGR system using all speech frames (Speech) and using only voiced speech frames (Voiced) for F0 and the cepstral features with 19 static and delta coefficients under three types of conditions: controlled, degraded and adverse for protocol **A**. *F0 for unvoiced frames was estimated using the Fourier interpolation method.

speech/non-speech segmentation was used to pick out speech frames (both voiced and unvoiced). In the second system, only voiced speech frames were selected. The latter system provided higher recognition rates for all the evaluated features, especially under noisy conditions. This is consistent with observations made in the previous studies where better AGR performance was reported on the voiced phonemes compared to unvoiced phonemes [3, 4]. In addition, our study showed that selection of voiced speech frames can make the gender recognizer robust towards mismatched training and testing conditions (protocol **A**). In the rest of the paper, we report results for the system employing the voiced speech frame selection.

4.2. Fundamental Frequency

We investigated suitability of F0 for the AGR problem under varying conditions and the obtained results are shown in Table 3. When the AGR system was trained exclusively on clean condition data (protocol **0** and **A**), F0 attained perfect recognition under controlled conditions, however recognition rate was highly affected by signal degradation. In order to analyze the results, distributions of F0 for females and males in the training set and in the three test sets (one for each conditions) are presented in Figure 1. The perfect recognition under controlled conditions is a consequence of almost an ideal match between F0 distributions for the training and test data. Under degraded conditions, the low frequency noise was also estimated as F0 resulting in the degradation of performance for females (91.7%) and perfect recognition for males. The adverse condition data were collected in noisy environment which presumably can force subjects to raise their F0 while increasing their voice intensities (*Lombard effect*). As a result, the distributions of F0 for both females and males were shifted towards high values, and the mismatch between data used for training and testing occurred. In this case, significant decrease in performance for males (86.1%) and perfect recognition for females were observed. The addition of noisy data to the training set (protocol **C**) decreased the performance under degraded conditions, since more incorrect estimates of F0, mostly from the degraded signal, were introduced to the system. On the other hand, the performance under adverse conditions increased, thus indicating that the system tries to compensate for the effect due to raised F0 values.

4.3. Cepstral Features

Figure 2 shows the results of protocol **0** (clean matched conditions) and protocol **A** (mismatched conditions). It can be observed that performance of all three cepstral features increases with the number of cepstral coefficients under all conditions.** This trend is consistent with the results obtained on high-quality

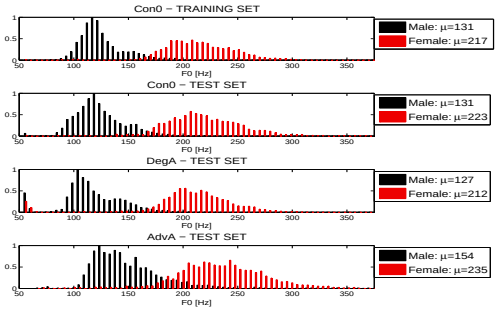


Figure 1: Distributions of F0 values for females and males in the training set containing controlled condition data and the three test sets consisting of controlled (Con_0), degraded (Deg_A) and adverse (Adv_A) data.

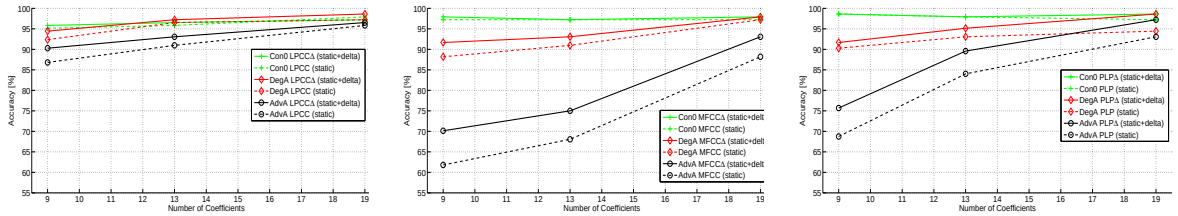
Feature	Accuracy [%]						
	Con_0	Deg_A	Adv_A	Deg_B	Adv_B	Deg_C	Adv_C
F0	100	95.8	93.1	95.8	93.1	95.1	94.4
$LPCC18_\Delta$	97.2	98.6	96.5	98.6	96.5	100	97.9
$MFCC19_\Delta$	97.9	97.9	93.1	97.9	93.1	98.6	99.3
$PLP19_\Delta$	98.6	98.6	97.2	98.6	97.2	97.9	98.6

Table 3: Performance of the AGR system for the voice source and vocal tract related features. Symbol Δ denotes use of both static and delta coefficients, e.g. for $PLP19_\Delta$ in total 19+19=38 coefficients were used. For LPCCs energy coefficient was not determined.

data by Wu *et al.* [3]. However, it is important to note that the increasing of number of cepstral coefficients aided in performances significantly more for degraded and adverse than controlled conditions. This leads to the conclusion that detailed modelling of spectrum is more crucial for noisy than clean-condition recordings. Second, the use of the delta coefficients in addition to the static coefficients further improved performance for all three cepstral features. This observation is consistent with the results obtained on clean-condition data by Fussell *et al.* [4]. Consequently, the system employing 19 static and delta coefficients under mismatch noisy conditions can almost approach the performance as under clean matched conditions. Furthermore, LPCCs are more stable features than MFCCs and PLPs, in the sense that the amount of degradation in performance due to reduction of number of cepstral coefficients is significantly lower for LPCCs than for MFCCs and PLPs. This is possibly owing to the differences in characterizing a smooth spectral envelope by these features. In case of LPCCs, spectral peaks of the short-term spectrum are modelled directly, whereas in case of MFCCs and PLPs, the spectrum resulting from human auditory related processing is represented. Additionally, what can be useful for AGR, the estimation of the spectral peaks based on the linear prediction is affected by F0 for voiced speech segments, since formant frequencies and bandwidths are sensitive to the value of F0. This needs further investigation and can be a part of the future work.

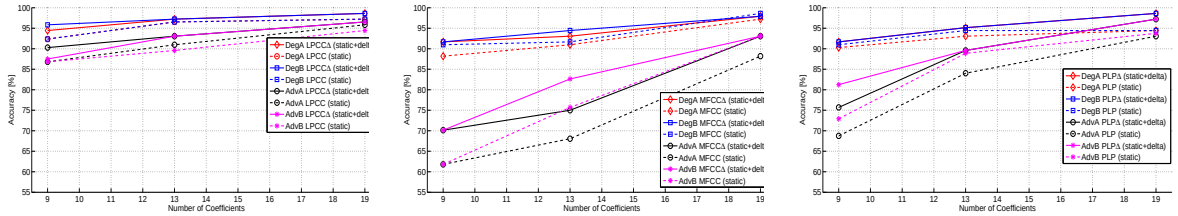
Figure 3 compares the results of protocol **B** (tuning of SVM parameters on noisy development data) with respect to the results of protocol **A** (mismatched conditions). The tuning of SVM parameters (γ, C) using development set specific for particular testing conditions slightly improved the performance of the cepstral features with lower number of coefficients.

Figure 4 compares the results of protocol **C** (multi-condition training) with respect to the results of protocol **A** (mismatched conditions). Not surprisingly, the performance of the system under noisy conditions improves with multi-condition training. It can be observed that with multi-condition training: (a) performance of the system is less dependent upon



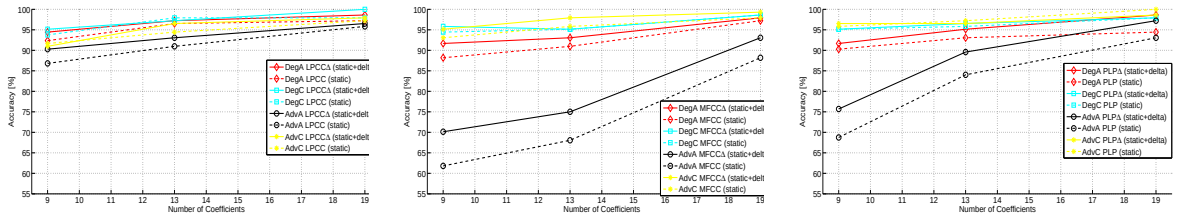
(a) Results for LPCCs. (b) Results for MFCCs. (c) Results for PLPs.

Figure 2: Performance of the AGR system for the cepstral features with respect to the number (9, 13, 19) and type (static vs. static+delta) of cepstral coefficients included to the feature vector under controlled (Con_0), degraded (Deg_A) and adverse (Adv_A) conditions. **For Con_0 slight improvements in performance with number of coefficients were observed for frame accuracy and flat characteristics in the figures are the consequence of presenting results for file accuracy.



(a) Results for LPCCs. (b) Results for MFCCs. (c) Results for PLPs.

Figure 3: Comparison of performance of the AGR system across protocols **A** and **B**, i.e. training and development on clean data vs. training on clean data and development on noisy data.



(a) Results for LPCCs. (b) Results for MFCCs. (c) Results for PLPs.

Figure 4: Comparison of performance of the AGR system across protocols **A** and **C**, i.e. training and development on clean data vs. training on clean+noisy data (multi-condition training) and development on noisy data.

the number of cepstral coefficients (i.e the amount of spectral details that are modelled); and (b) the system employing only static coefficients yields performance closer to the system using both static and delta coefficients. Thereby, the modelling of higher spectral detail and use of dynamic features is more important under mismatched conditions.

Table 3 summarizes the results for the cepstral features: LPCCs, MFCCs, and PLPs for the best setup (19 static and delta coefficients). It can be observed that in such setup, all the three cepstral features yielded comparable performance. Moreover, under clean conditions, the performance of F0 and the cepstral features were comparable. However, under noisy conditions, the cepstral features yielded more robust system. Also, multi-condition training helped more the cepstral-based system compared to the F0-based system.

5. Conclusions

In this paper, we analyzed F0 and cepstral features for robust automatic gender recognition. Through studies performed on BANCA corpus we showed that: (a) modelling only voiced speech frames improves the robustness of the AGR system towards mismatched conditions for both F0 and the cepstral features; (b) under clean matched conditions or with multi-condition training, the performance of the AGR system is less sensitive to the number of cepstral coefficients (i.e. the amount of spectral details being modelled); (c) modelling of higher spectral details and the use of dynamic features makes the system robust towards mismatched conditions; and (d) F0 and cepstral features provide similar performance under clean conditions, but cepstral features yields robust system in noisy conditions.

6. Acknowledgements

This work was supported by the EU 6th FWPISTIP AMIDA project (FP6-033812). The authors thank Sébastien Marcel and Jie Luo for help, discussions and comments.

7. References

[1] S. Bengio et al., "Experimental Protocol on the BANCA Database," Idiap, Tech. Rep., 2002.

[2] D. G. Childers and K. Wu, "Gender recognition from speech. Part II: Fine analysis," *JASA*, vol. 90, pp. 1841–1856, 1991.

[3] K. Wu and D. G. Childers, "Gender recognition from speech. Part I: Coarse analysis," *JASA*, vol. 90, pp. 1828–1840, 1991.

[4] J. W. Fussell, "Automatic sex identification from short segments of speech," in *Proc. ICASSP*, 1991, pp. 409–412.

[5] D. Talkin, *A robust algorithm for pitch tracking (RAPT)*, ser. Speech coding and synthesis. Elsevier Science, 1995.

[6] C.-C. Chang and C.-J. Lin, *LIBSVM*, 2001, software at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[7] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, 1999, pp. 61–74.