Gender Distinction Using Short Segments of Speech Signal

Milan Sigmund

Brno University of Technology, Purkynova 118, CZ-612 00 Brno, Czech Republic

Summary
This paper presents and discusses an approach to automatic gender distinction in a short segment of normally spoken continuous speech. In order to see which phonemes are effective for gender recognition, we analyzed individual vowels. Two different simple identifiers based on selected mel-frequency cepstral coefficients were evaluated. Using vowel phonemes, we achieved in short-time analysis (20 msec) a gender identification accuracy of more than 90%. Especially for vowel "a", almost no error occurs. For text-independent analysis, the speech duration of 500 msec was sufficient to identify male/female speakers with the accuracy of more than 93%. Automatic estimation of speaker's gender by her/his voice is an important factor to realize high-quality dialogue systems.

Key words:
Speech processing, gender recognition.

1. Introduction

Like all human behavior, speech is characterized by considerable variability. This derives from the plasticity of the speech apparatus, which is modified by a wide range of affective, stylistic, as well as internal and external environmental parameters. Each speaker has some associated properties, such as sex, age, dialect, profession, etc. Some control over these properties can be obtained by selecting effective test parameters or specific speech material.

Some studies in the literature show that speech recognition and speaker identification would be simpler, if we could automatically recognize a speaker's gender (sex). For example, in the "cocktail party effect", the voices of two or more speakers may be mixed. If the speakers are of opposite sex and if sex identification can be made on short segments of speech, the voices can be at least partially separated. Sex identification was used primarily as a means to improve recognition performance and to reduce the needed computation. Accurate sex identification has different uses in spoken language systems, where it can permit the synthesis module of a system to respond appropriately to an unknown speaker. In languages like French, where formalities are often used, the system acceptance may be easier if greetings such as "Bonjour Madame"are foreseen. In the past, automatic gender identification has been investigated for clean speech by Wu and Childers [1]. Clean speech and speech affected by adverse conditions are evaluated for a variety of gender identification schemes in [2]. Using speech segments with an average duration of 890 msec (after silence removal), the best mentioned accuracy is 98.5% averaged over all clean and adverse conditions.

There is some evidence that sex-related speech characteristics are only partly due to physiological and anatomical differences between the sexes; cultural factors and sex-role stereotypes also play an important part. Therefore, it is not well known at what age sex-related speech characteristics become prevalent. A comparison between male and female larynges on the basis of overall size, vocal fold membranous length, elastic properties of tissue and prephonatory glottal shape is drawn in [3].

Figure 1 illustrates the relationship between fundamental frequency of speech (i.e., DSP parameter) and membranous length (i.e., anatomical parameter). The fundamental frequency is scaled primarily according to the membranous length of the vocal folds, whereas mean airflow, sound power, glottal efficiency, and amplitude of vibration include another scale factor that relates to overall larynx size.

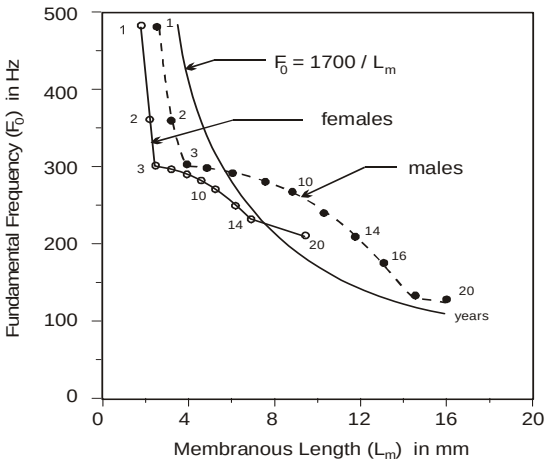

Fig. 1 Mean speaking fundamental frequency F_0 as a function of membranous length L_m (after [3]).

Manuscript received October 5, 2008
Manuscript revised October 20, 2008

The main feature which can speaker's sex distinguish is fundamental frequency F_0 with typical values of 110 Hz for male speech and 200 Hz for female speech. The pitch of children is so different that they are often treated as "the third sex". Most values of F_0 among people aged 20 to 70 years lie between 80-170 Hz for men, 150-260 Hz for women and 300-500 Hz for children [4]. There are Gaussian distributions of these ranges, so that dispersion is wide and we often could not categorize the acoustic signal reliably by using this criterion only.

There was predicted an inverse relationship between fundamental frequency F_0 and membranous length L_m with fixed tension and fixed mass per unit length.

In addition to the fundamental frequency, formants in some phonemes can also reliably distinguish speaker's sex. For example, the first three average formant frequencies F_1, F_2 and F_3 for vowels by adult female speakers have higher values than by male speakers [5].

2. Achieved Results

2.1 Experimental Speech Data

The speech data consisted of 420 sentences in total, 5 sentences by each of the 84 speakers (53 male and 31 female). All speakers in the database were subjective in good physical and psychical condition and have no speech, language or hearing difficulties. Most of the speakers are student aged 20 to 25 years. All speakers are Czech natives speaking with standard Moravian accent. The speakers were not informed of the objectives of the study before the experiment. The speech signal was sampled at 22 kHz using a 16-bit A/D converter under usual conditions in an office room.

The results obtained in previous research [6] indicate that vowels and nasals are generally the best phonemes for speaker identification. Thus, we focused our study on the sex-specific information contained in vowel phonemes.

2.2 Cepstral Analysis

The current most commonly used short-term measurements in speech signal processing are cepstral coefficients and their frequency-warped alternative coefficients.

Thus, the *mel*-frequency warped cepstral coefficients (MFCCs) were taken for our experiment to identify the sex of a speaker. First, a Hamming window was applied for each speech frame (20 msec) of the recorded vowels and the FFT spectrum was computed. Then, the spectrum was *mel*-warped and the inverse Fourier transform of the logarithm of the warped spectrum produced the vector of cepstral coefficients. The *mel*-frequency scale is linear below 1 kHz and logarithmic above 1 kHz [7].

Using a set of 40 *mel*-cepstral coefficients c_1 through c_{40} and their various differences, the performance of these individual features as identifiers of the sex of a speaker was measured. Table 1 summarizes the selected suitable coefficients which had the lowest variation calculated individually for all vowel phonemes and then averaged for both genders separately.

Table 1: Mean μ and standard deviation σ of selected *mel*-frequency cepstral coefficients.

MFCC	Male		Female	
	μ	σ	μ	σ
c_1	-326	122	-597	88
c_2	338	141	164	137
c_6	28	73	182	70
c_9	94	52	-20	53
c_{17}	21	65	129	45
c_{18}	31	39	112	47
c_{22}	20	46	109	57
c_{23}	-35	51	37	55
c_{24}	-1	39	98	29
c_{25}	5	57	158	28
c_{26}	-42	47	101	34
c_{35}	21	41	-119	80
c_{36}	60	45	-84	91

The best individual feature for sex identification seems to be the coefficient c_{24} followed by c_{26} and c_{25}, respectively. On the other hand, the differences of cepstral coefficient pairs are not reliable for sex identification [8].

2.3 Evaluated Gender Identifiers

Two sex recognition approaches were used in our test. The first approach was based on an individual cepstral coefficient. Applying an empirical formula to the coefficient c_{24} we get the gender identifier D_{24} in the form

$$D_{24} = |c_{24} - 80| - |c_{24} - 40| - 120 + 2c_{24} \quad (1)$$

This indicator gives a negative value for male and a positive value for female speakers.

The second approach used a set of selected cepstral coefficients according to the Tab.1. For both sex classes the reference mean vectors were formed as follows:

Male reference:

M = [-326, 338, 28, 94, 21, 31, 20, -35, -1, 5,…]

Female reference:

F = [-597, 164, 182, -20, 129, 112, 109, 37,…]

and the Euclidean distances d_1 and d_2 were calculated in each test

$$d_1(\mathbf{X},\mathbf{M}) = [(\mathbf{X}-\mathbf{M})^T (\mathbf{X}-\mathbf{M})]^{1/2} \quad (2)$$

$$d_2(\mathbf{X},\mathbf{F}) = [(\mathbf{X}-\mathbf{F})^T (\mathbf{X}-\mathbf{F})]^{1/2} \quad (3)$$

where **M** and **F** denote the reference vectors mentioned above, **X** is the tested vector formed by the same coefficients c_i as the reference vectors, and T denotes transpose. Computing the difference of the two distances

$$D = d_1(\mathbf{X}, \mathbf{M}) - d_2(\mathbf{X}, \mathbf{F}) \quad (4)$$

we get a measure which gives similar polarity result as the identifier D_{24} (negative for male, positive for female).

2.4 Gender Recognition Rate

Both procedures described above were evaluated for basic vowel phonemes, which provided an identification accuracy of more than 90%. Especially for vowel "a", almost no error occurs.

Table 2: Gender recognition rate in percent testing all vowels.

Identifier	Test Vowel				
	a	e	i	o	u
D_{24}	99	92	97	93	91
D	99	94	98	92	94

Table 2 shows the recognition rate obtained for all individual vowels cut out from a normally spoken speech. When the difference mel-cepstral coefficients were added the gender identification performance decreased slightly.

Both identifiers were also applied in a specific case, namely to test an imitation of female voice spoken by good male imitator. Using the vowel "a" this female-like voice was correctly recognized as male.

Although the sex identifiers were trained only on the vowel phonemes, we tried to apply the second identifier D to an entire fluent speech signal, i.e. to all phonemes. A sequence of values $D(j)$ corresponding to sex classification on each signal frame j gives the final result simply as a sum

$$\overline{D} = \sum_{j=1}^{J} D(j) \quad (5)$$

As expected, the voiced frames are more applicable than the unvoiced ones, so that a selector of voiced frames should be included before applying sex identification in this manner.

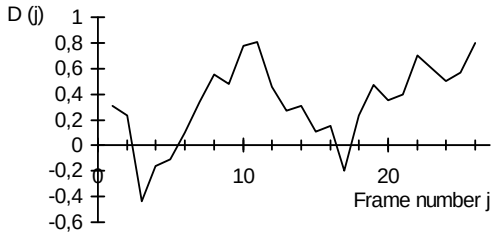

Fig. 2 Trajectories of sex identification values calculated for a female voiced speech.

Figure 2 illustrates scaled values of $D(j)$ for female voiced speech of 500 msec duration. In this case, some frames of consonants were represented incorrectly with inverse polarity of $D(j)$. For a long speech segment, the incorrectly identified speech frames are absorbed in the sum (5) by the majority of correctly identified frames. For a very short speech segment, the errors that randomly occurred on this segment can play a more important role.

The error rate in sex identification as a function of speech duration is shown in Fig. 3. Each speech segment used for the test was a part of a single sentence, preceded by a silence and always started at the beginning of the sentence.

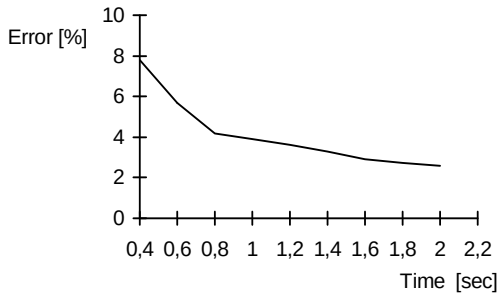

Fig. 3 Text-independent sex identification error rate as a function of signal duration.

The results in Fig. 3 show that the text-independent sex identification error rate is 6.8% with 0.5 sec of speech and 3.2% with 1.5 sec of speech. This implies that before the speaker has finished enunciating the first word, one is fairly certain of the speaker's sex. Speech duration of 0.5 sec seems to be sufficient for sex identification. To achieve a better performance of sex identification the terms of covariance matrix (diagonal or full) can be added to the decision rules.

3. Conclusions

We discussed gender recognition by voice and presented new algorithms for practical application. Some important conclusions about automatic gender distinction resulted from our research as follows:

- Using a set of selected *mel*-warped cepstral coefficients the sex of a speaker can be correctly identified with a performance of about 93% from speech segments with an average duration of 500 msec for clean speech.
- The time needed for sex identification decreases significantly for selected speech segments (e.g., vowel phonemes).
- The *mel*-frequency cepstral coefficient c_{24} seems to be the best individual feature for sex identification. Applying it to the phoneme "a", the sex identification was almost errorless.

In our future research we will try to test and modify the described gender identifiers for speech under various real-world conditions (noise, reverberation, coding speech for mobile telecommunication) and for other non-Slavic languages.

Acknowledgments

This work was supported by the Research Plan of Brno University of Technology MSM 0021630513 "Advanced Electronic Communication Systems and Technologies (ELCOM)" and by the project FRVS 1756/2008.

References

[1] K. Wu and D.G. Childers, Gender recognition from speech. Part I: Coerse analysis, *Journal of the Acoustic Society of America*, *90*(4), 1991, pp. 1828-1840.

[2] S. Slomka and S. Sridharan, Automatic gender identification under adverse conditions, Proc. *Eurospeech'97*, Rhodes, 1997, pp. 2307-2310.

[3] I.R. Titze, Physiologic and acoustic differences between male and female voices, *Journal of the Acoustic Society of America*, *85*(4), 1989, pp. 1699-1707.

[4] R.J. Baken and R.F. Orlikoff, *Clinical measurement of speech and voice*, Singular Publishing Group, San Diego, 2000.

[5] D. O'Shaughnessy, *Speech communication: Human and machine*, Addison-Wesley Publishing, Massachusetts, 1987.

[6] M. Sigmund, Effectiveness of various phonemes for speaker recognition, *Proc. Radioelektronika'01*, BUT, Brno, 2001, pp. 158-161.

[7] E. Zwicker and H. Fastl, *Psychoacoustics*, Springer Verlag, Berlin, New York, 1999.

[8] M. Kepesi, and M. Sigmund, Automatic recognition of gender, *Proc. Radioelektronika'98*, BUT, Brno, 1998, pp. 200-203.

Milan Sigmund received a masters degree in 1984 in biomedical engineering and a doctoral degree in 1990 in speech signal processing, both from the Brno University of Technology, Czech Republic. Currently, he is in the Faculty of Electrical Engineering and Communication at Brno University of Technology. In the years from 2001 to 2003, he stayed in the Department of Computer Science at the University of Applied Sciences Wiesbaden, Germany. His main research interests include speech signal processing with a special focus on automatic speaker recognition. He is a member of ISCA and EAEEIE.