ON THE ANALYSIS OF PHONEME-BASED FEATURES FOR GENDER IDENTIFICATION WITH NEURAL NETWORKS

Fikret S Gurgen[1], Ting Fan, Julie Vonwiller
Speech Technology Research Group
The Department of Electrical Engineering
The University of Sydney NSW 2006 Australia

ABSTRACT

As a first step in our research into the identification of non-linguistic features, we introduce the identification of speaker gender or sex using neural networks (NN). The study makes use of phoneme-based features (phonemes and broad phoneme classes) from a phonetically rich database, using a few of the sentences as training data, and investigates the effects on gender identification on a comparative basis. There were 3 speakers, 1 male and 2 females with different accents. Accurate identification of gender is known to increase the performance of speech recognition systems. Briefly, a window-based neural network (WNN) is generally trained for identification of non-linguistic features using phoneme samples. This NN is then used for testing for the feature in unknown phoneme samples. Various numbers of MFCCs are employed for the gender identification. It was found that vowels of just a few sentences provided valuable gender information.

INTRODUCTION

A major concern of current research in automatic speech recognition is the improvement of the accuracy and robustness of speech recognisers. One of the areas that may assist in this is the identification of "non-linguistic" speech features present in the acoustic signal. Non-linguistic features are aspects such as accent (Blackburn et al), language, dialect, speaker, gender, etc. It is generally believed that knowledge about these non-linguistic features would help improve the performance of recognition systems. For example, it is possible to envisage the application of accent identification where the spoken query is to be recognised without prior knowledge of the accent spoken. In societies consisting of people from different linguistic backgrounds, the performance of a speech recogniser is decreased due to the high degree of variation between speakers. Thus, the ability to automatically identify the accent being spoken and to adapt the recogniser appropriately would enhance the accuracy of the system.

Other applications of speech technology, such as voice access to a variety of computer and telephone-based services also require the usage of non-linguistic features. For example, at information centres in public places, such as bus stops, railway stations and airports, the language may change from one user to next. In order to become as user-friendly as possible, the information system should be able to recognise the spoken query without prior knowledge of the language being spoken so that the adaptation and the synthesis of speech for voice-responding system can satisfy the client.

Acoustic differences due to gender are useful non-linguistic information that help to improve the accuracy of speech (and also speaker) recognition systems. Gender identification can be viewed is a subset of speaker identification and assists the speech recognition and speaker verification or identification. In speech recognition, it is known that the use of sex-dependent models gives better recognition performance than speaker-independent models (Lamel et al). However, in large vocabulary systems this would mean that the recognition process would be carried out twice, once for each gender. A logical alternative is first to determine the speaker gender and then perform the recognition. For an improved recognition performance, a priori knowledge of gender is valuable.

Gender identification has been previously investigated using template matching with various feature parameters (Childers et al), Gaussian classifiers (Fussell) and ergodic hidden Markov models (HMM) (Lamel et al) which uses phoneme-based ergodic HMMs for the gender and trains it with a set of samples. Gender identification on the incoming signal is then performed by computing the acoustic likelihoods for all the models of a given set. The gender of the speaker is hypothesised as the gender

associated with the model set giving the highest likelihood.

Keeping the future needs of speech recognition in mind, this study presents an analytical approach that uses neural networks (NN) and phoneme-based features for identification of the gender of a speaker. The focus is to train the NN with phoneme-based feature parameters from multiple speakers with known genders. Then, the NN can have the ability to instantly distinguish the gender when the unknown input is applied to the network. The study also investigates the gender identification problem in a limited phoneme-based data trial using phonemes extracted from a full set of phonemes from several speakers of Australian English. For this purpose, it relies on linguistic knowledge to indicate the most useful features for gender distinction.

GENDER IDENTIFICATION WITH THE MFCC OF PHONEMES

The phoneme-based approach is used as the smallest set of distinguishing features of gender between speakers. Phonemes carry sufficient information for linguistic distinction of voices and act as a useful beginning point for investigation into non-linguistic features. The work of Lamel et al in this area also supports the effectiveness of phoneme-based features in identification of gender and other types of non-linguistic features. Furthermore, phonemes are generally used as the basis for large vocabulary speech recognition systems, and the additional use of the phonemes for identification of non linguistic features is an advantage. Thus, the compatibility between the elementary units of the gender identification and, for example, speech recognition improves the efficiency of the preprocessing stage. Fast and correct hypothesis of gender clearly affects the overall performance.

Mel frequency cepstral coefficients (MFCC) can be employed as a favourable choice of parameter set. By definition, MFCCs are inverse logarithmic Mel-scaled frequency parameters. In the vocal tract model (Furui), the output signal is obtained by the product of the vocal tract transfer function and the input source which can be either periodic pulses or white noise. These coefficients separate the effects of the vocal tract transfer function and the input source through logarithm operation. As a result, it may be advantageous to select them since they also, separately summarize the vocal system which is useful for identification of non-linguistic features.

The phonemes are derived from a read database of Australian English which consists of 200 phonetically rich sentences. Two strategies are used to approach the problem: The first strategy is to collect all specific phonemes such as /a/, /i/, /s/..., or some classes of phonemes, e.g. vowels & consonants, from the database and use them for the experiments. In this strategy, the phoneme information of the entire database is used for training of the NN classifier. The second strategy is to choose a number of sentences (1, 2, 3, 4) and use the phoneme information in just these sentences for training. Such a strategy makes possible the identification of gender of a speaker with a limited number of utterances.

As a first step, the phonemes of a short speech segment (for example, a sentence) are used as distinguishing features. In this case, after the speech signal is preprocessed, each segment can be used as input for the training and testing without consideration of which phoneme it is. This basic approach can be extended to incorporate some further linguistic knowledge. For example, the sound system can be divided into two broad classes: vowels and consonants and the effect of each class on the identification task can independently be examined. A further extension would be to select the phonemes representing the extremes of the vowel triangle as the gender marking features. The Australian English vowel triangle is given below in Figure 1. Here the peripheral vowels are /i a U/ and not /u/ which is more central in Australian English.

Using the broad phoneme classes as the potential gender identifying features, vowels were found to be effective in distinguishing genders, and more specifically, the vowels such as /a/ and /i/. These voweels have the added advantage of high frequency of occurrence.

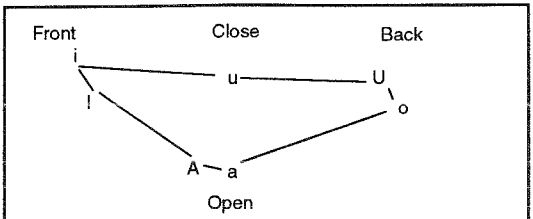

Figure 1. Some of the Australian English vowels at the periphery of the vowel triangle.

WINDOW-BASED NEURAL NETWORKS FOR GENDER IDENTIFICATION

This study employs NN-based classifiers (Lippman, Gurgen). The main aim is to extract phonemes from the unknown speech signal and to employ window structures of a WNN to search the distinctive features for gender discrimination. The WNN is a feedforward NN with window structure which is constructed with time-delayed frames. It uses a supervised training algorithm. The window structure seeks rather slowly changing features that are useful in recovering gender information. Windows of delayed frames (Figure 2) are shifted over the data features, and the features related to the distinction of gender are extracted using a time-invariant architecture.

A three layer WNN (Figure 2) is employed as an elementary architecture. The network uses one window (with 3 shifts) at the input layer to scan the gender features from an 8 frame fixed input size. In the second hidden layer, the information from the input layer is processed by another window (6 frames) which extracts the higher level features without any shifting operation over the data. The resulting information at the second hidden layer is slowly changing features of gender classification. At the output layer, a tied connection is used between 2 output units and all of the second hidden layer. All windows are tied connected to the succeeding layer.

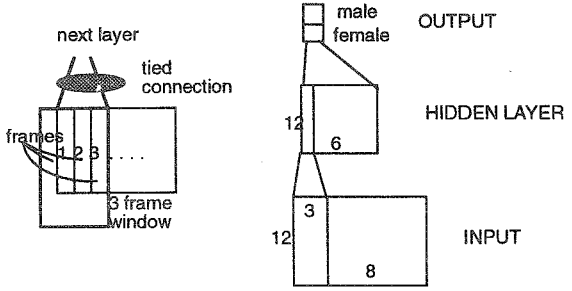

Figure 2 Window-based neural network (WNN)

The basic unit of the network evaluates the weighted sum of its inputs through a sigmoid function which is a continuous non-linear function. In this case, each sigmoid function in the unit receives the weighted sum of input values from delayed units.

A back propagation (BP) based recognition algorithm is used to train the networks. The training strategy gradually increases the number of samples used for training. The performance of the algorithm

improves by using variable momentum and learning rates during the computation of the new weight values. Convergence occurs roughly between 50-500 iterations and a maximum performance is obtained at the same range.

EXPERIMENTS AND RESULTS

In the experiments, data from 3 Australian English speakers was used- 1 male - broad Australian accent; 1 female - educated Australian accent; 1 female - general Australian accent. The materials were read speech from a database of 200 phonetically rich sentences. The sampling rate of the speech was 20 KHz and a 25 ms Hamming window was shifted over the data every 10 ms to obtain each frame; 12 order linear predictive coefficients (LPC) models were used to compute MFCCs as acoustic features of each phoneme; 8 fixed-size input frames were used for the WNN classifier. With these data, the study focused on (a) the effect of the number of coefficients on the identification accuracy and (b) the variation in the identification accuracy depending on the features analysed. Five experiments were performed where the features being examined for accuracy of gender identification were:-

1. vowel & consonant classes in 4 sentences, with varying number of MFCCs;
2. a single vowel /a/ using 10, 30 and 60 samples, with varying number of MFCCs;
3. two vowels /a/ and /i/ using 10, 30 and 60 samples;
4. four sentences with no specific phoneme information;
5. the vowel & consonant classes in increasing numbers of sentences.

Table 1: Training and test sentences used in the experiments

Training sentences	Test sentences
1. The catholic bishop said mass at the church	1. They were in a rush so they took a cab
2. He wanted recognition for his role in the jazz club	2. Tom would rather swim laps in the pool than go jogging
3. He paid a large cash deposit for the land	3. His tooth ached so badly he decided to visit the dentist
4. The corner of the roof had water dripping from it	

1. For the purpose of investigating the effect of the number of coefficients on the identification accuracy, the broad phoneme classes - vowels and consonants - were used in the 4 training sentences (See Table 2) with 4, 8 and 12 MFCCs for the identification. The vowel class provided the better gender classification.

Table 2: Accuracy of gender identification using vowel & consonant classes in 4 sentences with different numbers of coefficients (%)

Accuracy (%)	male1-female1			male1-female2		
Number of MFCC	4 MFCC	8 MFCC	12 MFCC	4 MFCC	8 MFCC	12 MFCC
Vowels	93.6	97.4	97.4	71.8	97.4	89.7
Consonants	68.6	74.6	69.5	64.4	81.4	83.9

2. Further generalization can be achieved by increasing the number of speakers and the number of sentences. The experiments were repeated using a single vowel /a/ from increasing numbers of sentences,

10, 30 and 60 for the training with 108 being used for the testing. (See Table 2

Table 3: Accuracy of gender identification using the phoneme /a/ in a larger sentence set, with different numbers of coefficients (%)

Accuracy (%)	distinguishing male1 & female1			distinguishing male1 & emale2		
Number of MFCC	4 MFCC	8 MFCC	12 MFCC	4 MFCC	8 MFCC	12 MFCC
Training data size(10)	89.8	96.3	90.7	83.3	90.7	89.8
Training data size (30)	92.6	100.0	98.1	86.1	96.3	92.6
Training data size (60)	97.2	99.1	100.0	89.8	97.2	92.6

The performance improved with the number of sentences used. It can be observed that the identification accuracy, in both cases, is closely affected by the number of the MFCCs. The performance is reduced by having fewer coefficients, and by having a large number of the coefficients. The optimal choice of 8 MFCC coefficients improved the performance and also reduced computation time. We continued the experiments with 12 MFCCs since our intention was also to investigate the effects of the various features on the identification and not to investigate the maximum achievable accuracy.

3. The /a/ and /i/ phonemes were collected from the database from respectively 10, 30 & 60 samples (and from 108 samples for testing). These were used as training data for gender identification (See Table 4). This experiment was designed to observe the effect of a single feature and sample size on the gender identification. It was observed that single features variably affected the accuracy of the identification. The performance was more variable but in general the greater the number of training samples the better the discrimination

Table 4: Accuracy of gender identification using /a/ and /i/ from 10, 30 and 60 samples (%)

Accuracy (%)	/a/			/i/		
Training data size (/a/ and /i/)	10	30	60	10	30	60
distinguishing male1-female1	90.7	98.1	100.0	89.9	96.3	92.6
distinguishing male1-female2	87.0	92.6	92.6	93.5	94.4	96.3

4. All the phonemes in 1 (2, 3, 4) sentences are selected for training and another 3 sentences are chosen for testing. The accuracy of identification is as shown in Table 5. It was observed that a few training sentences unspecified for phonemes were sufficient for the gender information but without the knowledge of phonemic classes, the accuracy did not improve as the number of sentences increased

Table 5: Accuracy of gender identification using sentences alone with an increasing no. of sentences (

Number of sentences/ accuracy (%)	1	2	3	4
male1-female1	74.0	80.1	79.1	79.6
male1-female2	78.1	82.1	86.2	85.2

5. Vowel and consonant classes were extracted from the sentences 1, (1 and 2), (1 and 2 and 3), and (1 and 2 and 3 and 4) for the identification of the same speaker set. The results of the discrimination using one, two, three or four sentences is compared in Table 6. It was found that the broad classification of vowels demonstrated increased accuracy as the number of sentences increased. Consonant accuracy increased over three sentences.

Table 6: Accuracy of gender identification using vowel & consonant classes in increasing numbers of sentences (%)

Number of sentences/ accuracy(%)	1	2	3	4
female1-male vowels	84.6	89.7	92.3	97.4
female1-male consonants	57.6	64.4	72.9	69.5
female2-male vowels	73.1	82.1	83.3	89.7
female2-male consonants	78.8	80.5	85.6	83.9

CONCLUSIONS

The study investigated gender identification performance of Australian English speakers using phoneme-based MFCC features by WNN structure. It utilized various numbers of MFCCs, phonemes, and broad phoneme-classes, from a few sentences and from the whole database. In the experiments with a limited numbers of speakers, the gender information can be derived at nearly same accuracy from both the few sentences or the whole database, but employing a limited number of sentences is more practical. Furthermore, broad classification into vowels and consonants, which is easily obtained from the speech, becomes valuable. Vowels provide more characterising information for gender discrimination than consonants. The experiments performed here on gender identification can be extended to other non-linguistic features such as accent, language, dialect, and speaker. In future work, we plan to continue to extend our research into the identification of non-linguistic features by increasing the number of speakers for gender identification and to extend the work to identify the accent of the speaker from his/her speech.

NOTES

1. Gurgen is on leave from Computer Engineering Department, Bogazici University. Bebek 80815 Istanbul, Turkey.

REFERENCES

Blackburn C , Vonwiller J , King R (1993) . "Automatic accent classification using artificial neural networks," Eurospeech '93, pp. 1241-1244.
Childers D , Wu K, Bae K , Hicks D. (1988) "Automatic recognition of gender by voice," ICASSP-88.
Furui, S. (1989) "Digital Signal Processing, Synthesis and Recognition", Marcel Decker 1989.
Fussell J. (1991) "Automatic sex identification from short segments of speech," ICASSP '91.
Lamel L , Gauvain J , (1993) "Identifying non-linguistic speech features," EuroSpeech '93 pp. 23-30
Lippmann R. (1991) "Review of Neural Networks for Speech Recognition," Reading in Speech Recognition (eds) Alex Waibel & Kai Fu Lee, Morgan Kaufmann Publishers, Inc. pp. 374-392, 1991.
Hush D. & Horne B, (1993) "Progress in Supervised Neural Networks," IEEE Sig. Proc. Mag., pp. 8-39
Gurgen F, Aikawa K, Shikano K, (1991) "Phoneme recognition with neural networks using a novel fuzzy training algorithm," IEEE IJCNN '91 pp. 572-577