Improving speech recognition performance through gender separation

W. H. Abdulla & N. K. Kasabov
Information Science Dept./ University of Otago
Dunedin – New Zealand

Abstract

Speaker attributed variability are undesirable in speaker independent speech recognition systems. The gender of the speaker is one of the influential sources of this variability. Common speech recognition systems tuned to the ensemble statistics over many speakers to compensate the inherent variability of speech signal. In this paper we will separate the datasets based on the gender to build gender dependent hidden Markov model for each word. The gender separation criterion is the average pitch frequency of the speaker. Experimental evaluation shows significant improvement in word recognition accuracy over the gender independent method with a slight increase in the processing computation.

1. Introduction

Most parametric representations of speech are highly speaker dependent, and probability distributions suitable for a certain speaker may not be suitable for other speakers. Examples of speaker-dependent parameters are age group, differences in regional accents and the length of male and female vocal tracts. In common approaches of training speaker independent models these speaker dependent parameters are not considered during training. These recognition systems are instead tuned to the ensemble statistics over many speakers [1, 2].
It is better to put some consistency constraints to tag the utterance to the exact group that most likely has verbalised it. Given enough training data separate models, for each word, for male and female speakers can be trained to improve the recognition performance. For example, the performance of the SPHINX-II[♣] ASR system improved from adding gender-dependent parameters
We have now two options to proceed in the recognition phase. Either we use the whole models for the recognition, which is more accurate, but time consuming as we double the number of models to be aligned to the spoken word. Or during the recognition phase the utterance is categorised first to male or female group then the only models' set belonging to the decided category is used for recognition. The categorisation can be done based on the average VQ distortions [3, 4]. Suppose $\{f_k^{gender}\}$ is the k^{th} codeword density function for a specific gender, male or female. Given a segment of speech X, the likelihood of a word produced by the given gender can be approximated by

$$L(X \mid gender) = \prod_{x \in X} \sum_{k \in \eta(X)} f_k^{gender}(x) \qquad \ldots(1)$$

where x is one component frame of segment X and $\eta(x)$ contains the few codewords that best fit the acoustic frame x. If L(X|male) > L(X|female), the input speech is assumed to be produced by a male speaker, otherwise it is from a female speaker. One thing to note here is that the background silence must be excluded from X before processing.
After the gender is determined, only the models of the determined gender are activated for the recognition process.
The other possibility of gender determination is by using a multilayer neural network trained properly under supervision on the two genders [5].
In our system it was found that the average pitch frequency of the speaker is a simple and strong cue for categorising the genders as will be described in the next section.
This method showed that if the average value of the fundamental frequency is less than a certain threshold then the speaker is male otherwise the speaker is either kid or female. Figure (1) shows the block diagram of the gender models separation operation.
This paper is organised as follows:
In section 3 we will describe the methodology used in categorising the gender of the speakers according to their average pitch frequency. Section 4 depicts the experimental results and the evaluation procedure to assess the gender separation technique. Then we will derive the conclusions in section 5.

[♣] SPHINX is an ASR system developed by Carnegie Mellon University.

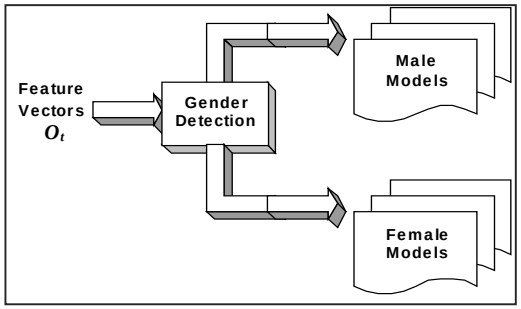

Figure (1) Gender detection step before feeding to the HMM recognition procedure.

2. Pitch category detection

Pitch frequency or fundamental frequency is defined as the frequency at which the vocal cords vibrate during voiced segments of the speech signal. It is an important parameter in speech analysis, synthesis, and coding. It is less valuable in speech recognition applications. It can be used in discriminating voiced from unvoiced speech frames, which is considered as a feature in some speech recognition systems. In this thesis we will use it differently and more effectively in the determination of the gender of speakers by exploiting the property of pitch period difference between genders. The developed speech recognition system models the utterances from different genders by two different models for the same utterance since it is more accurate than the mixed gender modelling.

However, accurate and reliable measurement of the pitch period is very difficult due to several reasons:

1 – The glottal excitation waveform is not a perfect train of periodic pulses.

2 – The pitch period varies even with the same speaker, speaking the same word in different emotional states.

3 – The pitch period is influenced by the accent of the speaker.

4 – There is an interaction between the pitch frequency and the formant frequency, resonance frequency of the vocal tract, which makes it difficult in some cases to differentiate between them.

5 – Pitch detection is sensitive to the environment and channel changes.

Due to the importance of the pitch detection, a wide variety of techniques have been proposed in the speech processing literatures [6, 7]. A comparative performance study between some classical techniques can be found in [8]. Other more successful pitch detectors based on more developed techniques such as wavelets and perceptual correlograms are showing more robust characteristics against noise and phase changes [9, 10].

Fortunately, the pitch detection for gender discrimination is not required to be very precise since we are interested in the frequency band that the fundamental frequency lying in, rather than the exact pitch frequency. The adopted pitch detection technique is based on the cepstral algorithms. Each frame of 40ms is weighted by equal length hanning window and the cepstrum of that frame is computed. The peak cepstral value and its location are determined. If the peak exceeds certain level the frame is assigned as voiced segment otherwise it is unvoiced one, and the pitch period is determined from the location of that peak. Figure (2) shows the real cepstrum of a voiced segment of 4 pitch periods and the effect of widowing on the peaks. Without windowing parasitic leakage peaks, from the harmonics due to the sharp ends of the segment, might interfere the pitch detection.

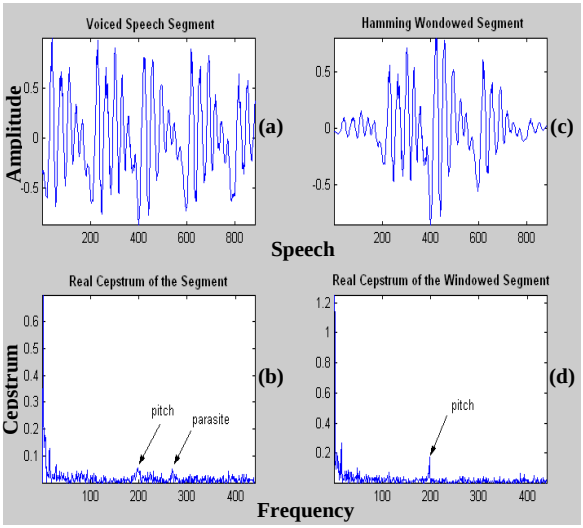

Figure (2) Cepstrum analysis of the voiced speech segment (a) voiced speech segment, (b) real cepstrum without windowing, (c) windowed speech segment, (d) real cepstrum of the windowed speech segment.

A post processing stage is added for error correction and to perform some smoothing on the pitch contour. Error checking is made to assure that an unvoiced frame cannot occur between two voiced frames. If this case appears the offending frame is reassigned again as voiced and its pitch value is the average value of the surrounding voiced frames. The smoothing can be achieved by using simple type of filters known as median of N filter. This filter operates on N input samples x(n), x(n-1), ..., x(n-N+1), places them in ascending order of amplitudes and selects the median as the filter output. Filter length N

of 3 has been used in the developed system and found useful in correcting occasional errors.
Figure (3) shows the pitch detection process of the spoken word five by a male speaker. The first and the last fricatives of the spoken digit five as well as the silence periods have zero fundamental frequency.
The discrimination index is the average fundamental frequency during the voiced periods.

$$\text{aveF}_0 = \frac{1}{N_v}\sum_{i=1}^{N_v} F_0(i) \qquad ...(2)$$

where $F_0(i)$ is the fundamental frequency (1/pitch period) of the i^{th} frame and N_v is the total number of the voiced speech frames.
A similar technique was first used to adaptively change the analysis frame size for pitch detection based on the autocorrelation method [11].
The average value of the fundamental frequency, aveF0, for this example was calculated to be 133Hz. This method showed that if the average value of the fundamental frequency is less than 160Hz then the speaker is likely to be a male, otherwise the speaker is either a kid or a female. This method was tested with TIMIT speech corpus and Otago speech corpus♠ and showed 100% gender discrimination accuracy. It was also tested with spontaneous real time input speech and some errors occurred only during attempts to imitate the speech of the other gender. Mostly when males imitated females and less with the reverse tries.

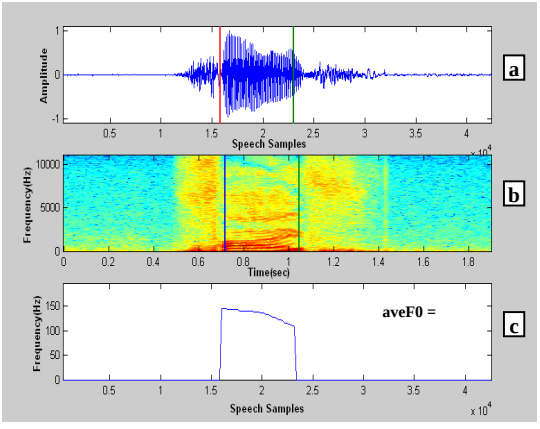

Figure (3) Fundamental frequency determination of a spoken digit five. The voiced segment is shown between the two lines. (a) time signal, (b) spectrum of the signal, (c) smoothed contour of the fundamental frequency.

♠ Otago speech corpus can be downloaded freely from http://Kel.otago.ac.nz/hyspeech/corpus

3. Experimental results

To have an idea about the performance improvement due to the gender inclusion we have done the followings:
1 – A model, λ_P, of a certain word "letter" was constructed using pooled examples of both male and female speakers. The model was 9 states CDHMM using 39 classical MFCCs.
2 – Another two similarly constructed models of the word "letter" were also constructed; one based on male speakers' examples, λ_M, and the other based on female speakers' examples, λ_F.
As a comparison of the state clustering of the three prepared models the Gaussian ellipsoids are plotted in figure (4). This figure illustrates that each state of the λ_P model is approximately averaging its corresponding two states of the models λ_M and λ_F.

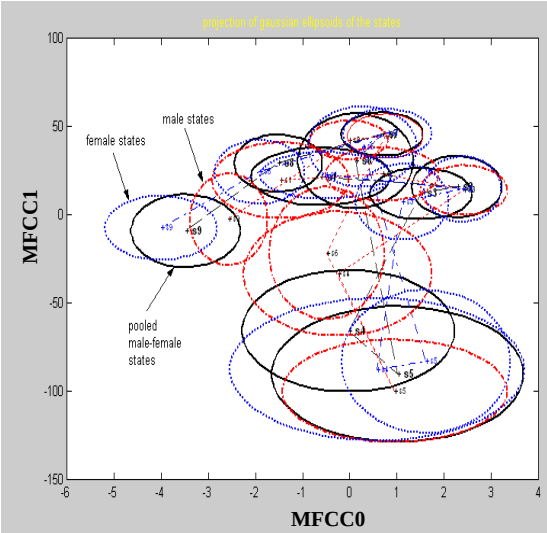

Figure (4) Gaussian ellipsoids of the three CDHMM, unimixture, 9 states models λ_P, λ_M and λ_F of the spoken word "letter".

3 – Several test examples of the same word "letter" were then presented to the above three models and the log-likelihoods, $P(O|\lambda_I)$, were calculated, where the subscript I refers to P, M, or F. The log-likelihoods of the presented examples from male speakers to the three models versus the average pitch frequency are plotted in figure (5).
It is as expected that the highest $P(O|\lambda_I)$ is that produced by the male speakers' model, $P(O|\lambda_M)$. The pooled model, $P(O|\lambda_P)$, is always contributed to the second highest log-likelihood, while the least is coming from the female speakers' model, $P(O|\lambda_F)$.
A similar procedure had been followed to produce the log-likelihoods of the same models as subjected to observations from

female speakers. The highest registered log-likelihoods in this case were coming from the female speakers' model λ_F as illustrated in figure (6).

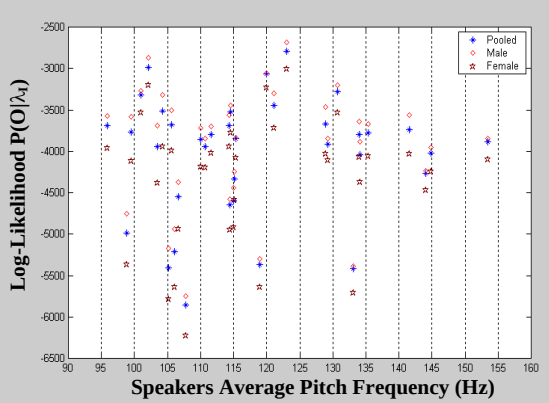

Figure (5) Comparison of the log-likelihoods produced by the three models λ_P, λ_M and λ_F when subjected to observations from male speakers.

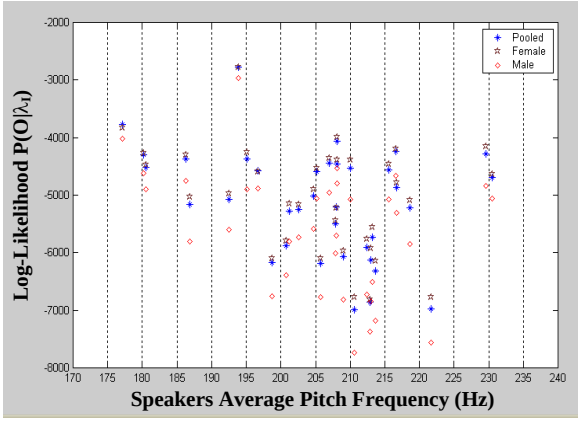

Figure (6) Comparison of the log-likelihoods produced by the three models λ_P, λ_M and λ_F when subjected to observations from female speakers.

From figure (5) and figure (6) we can see that the log-likelihoods produced by λ_P is somewhere between the two extremes of λ_M and λ_F and nearest to the best performing one. This mediation behaviour of λ_P is in consistency with the results depicted in figure (4).

The average $P(O|\lambda_I)$ produced by the three models as subjected to several examples from both male and female speakers were calculated and depicted in Table 1. This table shows that the maximum log-likelihood is the one that belongs to the matched word-gender-model examples, differentiated by two brackets. However, on the average if we have pooled male-female examples the best performance is of the model prepared from pooled examples.

	Average (x 10^3)		
	$P(O\|\lambda_P)$	$P(O\|\lambda_M)$	$P(O\|\lambda_F)$
Male examples	-4.0457	$(-3.9236)_{max}$	-4.3365
Female Examples	-5.1194	-5.6432	$(-5.0271)_{max}$
Pooled Examples	$(-4.5826)_{max}$	-4.7834	-4.6818

Table 1: Average log-likelihood of $P(O|\lambda_P)$, $P(O|\lambda_M)$ and $P(O|\lambda_F)$ as recorded by different examples of the word "letter".

4 – A set of different words was presented to the same models and the log-likelihood of the k^{th} observation sequence O_k is normalised by its length, N_k, according to the formula.

$$P_{norm}(O_k|\lambda_I) = \frac{P(O_k|\lambda_I)}{N_k} \qquad \ldots (3)$$

The results showed that the highest log-likelihood was produced when there was a matched word-gender-model example, as indicated in figure (7).

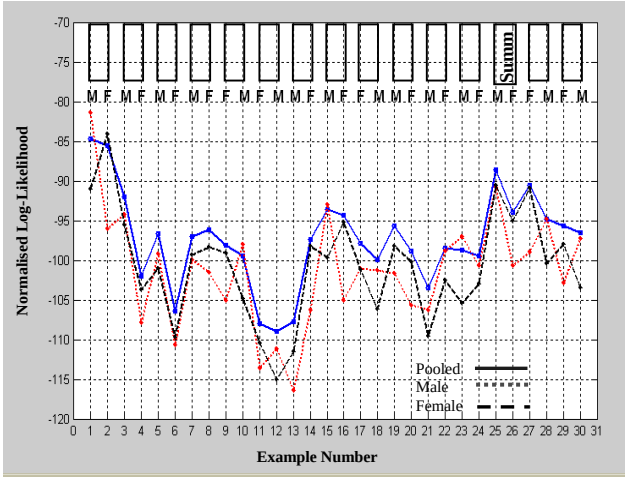

Figure (7) Performance evaluation of the three models λ_P, λ_M and λ_F of the word "letter" as subjected to observation from 15 different words spoken by male and female speakers. The spoken words are shown on the top, M and F refer to male and female speakers respectively.

The highest $P(O|\lambda_I)$ is that calculated from presenting the word "letter" spoken by a male and a female speakers to their corresponding models λ_M and λ_F. This is logical as they are matched word-gender-models. The other thing that can be derived from this figure is that in most cases $P(O|\lambda_P)$ is no longer be in between $P(O|\lambda_M)$ and $P(O|\lambda_F)$ for the words other than

"letter". Also, for the acoustically similarly words to the word "letter" the female speakers stimulate the female models more than the male speakers and vice versa which implicate the high information carried by the speaker related to the recognition process. These conclusions support the use of the gender specific models.
The recognition rate of a system based on gender dependent modelling is calculated. It shows 2% and 5% improvement in recognition rate when tested on datasets composed of 30 and 54 words respectively.

4. Conclusions

Gender dependent modelling role in improving the recognition performance has been studied and analysed through out several experiments. We have seen that the best performance is achieved when we have matched word-gender-model situations. We used the average pitch frequency of the speaker as a discriminating factor to identify the gender. A reliable method to find the average pitch frequency has been described. The precise pitch frequency is not required in our approach as we are interested in category rather than specific values. This method was tested with TIMIT continuous speech corpus and KEL isolate words speech corpus♠ and showed 100% gender discrimination accuracy (no error recorded).

5. References

1. Huang, X.D. *Minimizing speaker variation effects for speaker-independent speech recognition*. in *Proceedings of Speech and Natural Language Workshop*. 1992.
2. Hwang, M.Y. and X.D. Huang. *Subphonetic modeling with Markov states-senone*. in *Proc. IEEE ICASSP'92*. 1992.
3. Huang, X.D., et al. *Improved acoustic modeling for the SPHINX speech recognition system*. in *Proc. IEEE ICASSP'91*. 1991. Toronto, Canada.
4. Hwang, M.Y., *Subphonetic acoustic modeling for speaker independent continuous speech recognition*. 1993, CMU.
5. Abrash, V., et al., *Connectionist gender adaptation in hybrid neural network/hidden Markov model speech recognition system.* Proc. ICSLP'92, 1992.
6. Papamichalis, P., *Practical approaches to speech coding*. 1987, New Jersey, USA: Prentice-Hall, Inc., Englewood Cliffs.
7. Hess, W., *Pitch determination of speech signals*. 1983, New York: Springer-Verlag.
8. Rabiner, L., M.J. Cheng, and A.E. Rosenberg, *Acomparative performance study of several pitch detection algorithms.* IEEE Trans. ASSP, 1976. **24**(5): p. 399 - 417.
9. Slaney, M. and R.F. Lyon. *A perceptual pitch detector*. in *ICASSP'1990*. 1990.
10. Kadambe, S. and G.F. Boudreaux-Bartels, *Application of the wavelet transform for pitch detection of speech signals.* IEEE Trans. Information Theory, 1992. **38**(2): p. 917 - 924.
11. Rabiner, L., *On the use of autocorrelation analysis for pitch detection.* IEEE Trans. ASSP, 1977. **25**(1): p. 24 - 33.

♠ KEL speech corpus can be downloaded from http://Kel.otago.ac.nz/hyspeech /corpus