ROBUST GENDER-DEPENDENT ACOUSTIC-PHONETIC MODELLING IN CONTINUOUS SPEECH RECOGNITION BASED ON A NEW AUTOMATIC MALE/FEMALE CLASSIFICATION

Rivarol Vergin *Azarshid Farhat* *Douglas O'Shaughnessy*
INRS-Télécommunications
16 Place du Commerce, Île-des-Sœurs, H3E1H6, Québec, Canada
email: farhat@inrs-telecom.uquebec.ca

ABSTRACT

In this paper we present a new automatic male/female classification method based on the location in the frequency domain of the first 2 formants. This classification is based on a new automatic formant extraction which is faster than a peak picking technique. Gender-dependent acoustic-phonetic models stemming from this classification are used in the INRS Continuous speech recognition system with ATIS corpora. An improvement of 14% is obtained with these models in comparison to the baseline speaker-independent system.

1. INTRODUCTION

Today, most of the research in speech recognition is focused on the speaker-independent speech recognition problem. Generally the parameterization techniques used for speaker-dependent or independent recognition is the same. However, in the case of speaker-independent speech recognition, it is well known that the performance of recognizers for female speakers is almost always worse than that obtained for male speakers. The common solution adopted by several researchers is to have separate male and female acoustic-phonetic models. The training corpora is split up -manually or automatically- into female and male speakers.

Generally the automatic male/female classification is achieved based on the average value of the fundamental frequency, F0. However the distinction between men and women could also be represented by the location in the frequency domain of the first 3 formants for vowels [1], since we know that men and women have different formant positions for vowels. Our automatic male/female classification is based on the difference of position of the first and second formants between men and women; when we add the position of the third formant, the performance of our classifier does not increase.

The basic idea is to detect the gender of each speaker (in both training and testing corpora) with a robust but fast algorithm. We use a new automatic formant extraction technique which performs a detection of energy concentration. This algorithm provides good results and is much faster than the classic peak picking technique [2].

In the section 2 we describe our new formant extraction algorithm based on an iterative and fast research of the energy concentration in successive intervals. In section 3 we describe the automatic male/female detection which is based on the location of the first and second formant. Sections 4 and 5 provide the description of the INRS continuous-speech recognizer, the experimentation conditions and comparative results obtained by gender-dependent acoustic phonetic modelling.

2. FORMANT EXTRACTION

The new algorithm of formant-position evaluation presented in this paper is built on two main principles: (a) each formant has an associated bandwidth, (b) each formant falls in a known interval. The principle (a) allows us to look for a position in the spectra where there is a reasonable concentration of energy, and principle (b) allows us to delimit for each formant (first and second) the area in the frequency domain where we have to look for this concentration of energy. The process described in this paper reduces significantly the interval where the first or second formant can be found.

Let us define Y as the spectral energy vector obtained from the fast Fourier transform of the pre-emphasized signal:

$$y(n) = x(n) - \alpha x(n-1),$$

where $x(n)$ is the input signal and α is approximately equal to 0.95.

Assuming that the formant we are looking for lies between frequency positions k_b and k_e, the algorithm allows us to increase k_b or to decrease k_e by a fixed amount Δk until the interval $I = k_e - k_b$ reachs a predefined value I_{stop}. It is described by the following steps:

1. Calculate E_1:

$$E_1 = \sum_{k=k_b}^{k_b-1+I/2} Y^*(k)$$

2. Calculate E_2:

$$E_2 = \sum_{k=k_b+I/2}^{k_e} Y^*(k)$$

3. If $E_1 > E_2$ then $k_e = k_e - \Delta k$; else $k_b = k_b + \Delta k$.

4. Calculate I:

$$I = k_e - k_b$$

5. If $I > I_{\text{stop}}$ then return to step 1; else terminate.

The vector Y^* used in the algorithm will be defined later. An illustration of the process is shown in figure 1. Since the goal is not to find an exact position of the formant but an estimate good enough to allow us to discriminate between men and women, a reasonable estimate is:

$$P_{est} = \frac{1}{A} \sum_{k=k_b}^{k_e} kY^*(k),$$

where:

$$A = \sum_{k=k_b}^{k_e} Y^*(k).$$

P_{est} is simply the mean position of the energy in the interval $[k_b, k_e]$. Assuming that the center of the formant lies in each new interval given by the algorithm, when the final interval tends to zero, P_{est} tends to the true value of the center formant position.

In our implementation, the speech signal is sampled at 16 kHz. A spectral energy vector Y of 256 elements is evaluated every 10 ms using a Hanning window of 30 ms. Each element $Y(k)$ has a frequency resolution of 31.25 Hz. Δk has been fixed at 4 (125 Hz) and I_{stop} has been fixed at 8 (250 Hz).

2.1. First formant estimation

To estimate the position of the first formant, we need an initial interval valid for men and women. k_b has been fixed at 4 (125 Hz) and k_e at 28 (875 Hz). These values are deduced from table 1 where the average value of the first formant (men women) for ten vowels are provided by Peterson and Barney [1]. For the first formant evaluation, Y^* is taken as equal to Y.

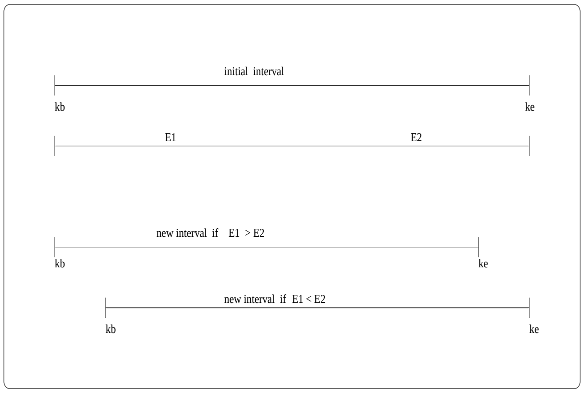


Figure 1. Illustration of the procedure of obtaining a reduction of the initial interval.

2.2. Second formant estimation

To estimate the position of the second formant, an initial interval is needed as in the case of the first formant. k_b is taken equal to:

$$k_b = \max(k_{F1} + 8, 28),$$

where k_{F1} is the estimated position of the first formant, and k_e has been fixed at 92 (2875 Hz) because the second formant is rarely beyond this value.

The spectral vector Y^* used to evaluate the second formant is a weighted version of Y. This is made through the use of a balancing vector γ defined by:

$$\gamma(k) = exp(\alpha * |k - 48|),$$

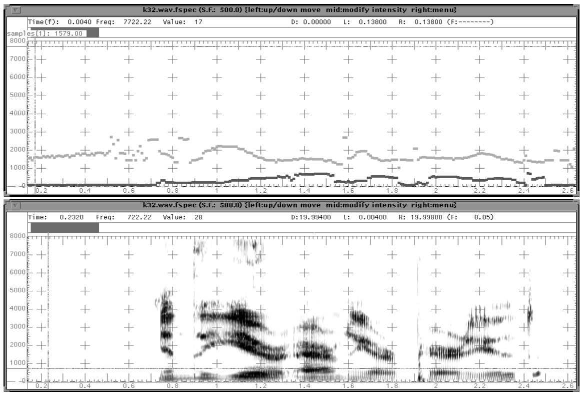
with $\alpha = log(0.1)/208$. The weighted spectral energy vector is given by:

$$Y^*(k) = (1.0 + Y(k))^{\gamma(k)}$$

The weighting vector γ highlights the energy around 1500 Hz which compensates for the high concentration of energy often observed in the first formant region. An example of the first two formants as found by the algorithm is shown in figure 2.

3. AUTOMATIC MALE/FEMALE CLASSIFICATION

As we pointed out before, the location of the first three formants is very different for men and women. Table 1 provides the average formant frequencies for English vowels by adult male and female speakers. We use the values of the first 2 formants as a reference for our classification. It



```
k32.wav.fspec (S.F.:  500.0) {left:up/down move  mid:modify intensity  right:menu}
Time(f):  0.0040 Freq:  7722.22   Value:  17              D: 0.00000   L:  0.13800   R:  0.13800 (F:--------)
samples[1]: 1579.00
```

```
k32.wav.fspec (S.F.:  500.0) {left:up/down move  mid:modify intensity  right:menu}
Time:    0.2320    Freq:    722.22   Value:  28           D:19.99400   L:  0.00400   R: 19.99800 (F:     0.05)
```

Figure 2. The upper part shows the first two formants of the sentence "*in the year earlier period*" and the lower part shows its spectrogram.

should be pointed out, that using the third formant does not bring an improvement in our classification results.

The classification is achieved on the two first sentences pronounced by each speaker (an average of 7 seconds per sentence). The consonant frames are detected by a comparison of the energy concentration in low and high frequency. The silence is determined using a minimum threshold of the energy.

For each speaker we compute 2 scores corresponding to the number of times the formant positions of a frame are assigned the male (and respectively the female) values. To do this, the formant locations of the vowel frames are compared with the reference male/female formant locations of all vowels. The least difference provides the gender associated to this frame. The corresponding score (male score or female score) is increased by 1. At the end of this computation, the greater score determines the estimated gender of the speaker.

4. INRS SPEECH RECOGNIZER OVERVIEW

In the INRS speech recognizer [3], speech data are divided into temporal blocks. This block processing brings an efficient solution to both real-time issues and memory problems associated with the classical forward-backward algorithm.

The recognition process is based on a two-pass search technique on the current block. The results of each block are passed to the next one for a new two-pass search until no more data is available.

The first pass produces a word graph using coarse acoustic and language models. This word graph is rescored in the second pass using fine acoustic and language models.

	F1		F2		F3	
	male	female	male	female	male	female
/i/	270	310	2290	2790	3010	3310
/I/	390	430	1990	2480	2550	3070
/ε/	530	610	1840	2330	2480	2990
/æ/	660	860	1720	2050	2410	2850
/α/	730	850	1090	1220	2440	2810
/⊃/	570	590	840	920	2410	2710
/U/	440	470	1020	1160	2240	2680
/u/	300	370	870	950	2240	2670
/∧/	640	760	1190	1400	2390	2780
/3/	490	500	1350	1640	1690	1960

Table 1. Average formant frequencies (in Hz) for English vowels by adult male and female speakers. (After Peterson and Barney [1]).

The acoustic-phonetic models are three-state left-to-right HMMs with no skip transitions. At the present time these HMMs represent right context-dependent phones in both the first and second passes. The output distributions of the HMMs were modelled by tied-mixtures.

We used statistical language models (n-grams) in our system. The coarse language model used in the first pass is a bigram model. The fine language model used in the second pass is a trigram model. To cope with the "*unseen events*" problem, the interpolated estimation method is used.

5. EXPERIMENTAL CONDITIONS

5.1. Automatic gender detection

First, the automatic gender classification is performed on all speakers of the training corpora, which are split up into male and female classes. Each set of speakers is used to estimate the corresponding gender-dependent acoustic-phonetic models. In the testing phase, the gender of each speaker is determined by the classification process using the first 2 sentences pronounced by the speaker. Then the corresponding acoustic-phonetic models are used in the recognition phase.

In both training and testing sets, 15% of speakers are wrongly classified. A classification is considered to be an error when the classifier finds a male speaker, whereas listening to the sentence seems to indicate a female (or vice versa). However we do not correct this distribution of the training (or the testing) corpora. The sentence, based on fundamental frequency F0, can appear to come from a woman while the formant positions are closer to formant locations for men.

5.2. Comparative Results

The speech corpus used in these experiments came from ATIS (Air Travel Information System) corpora, with a vocabulary of 1087 words. 285 speakers with a total of 9269 sentences are used for the training. The tests are performed with 10 other speakers for a total of 201 sentences. Males and females are present in both training and testing sets. In our baseline system (speaker independent case) we obtained 88.42% in word accuracy on the testing set. With our new male/female classification, this performance is increased to 90%.

	W. A. (%) (male)	W. A. (%) (Female)	W. A. (%) (Average)
1)	90.6	85.7	88.4
2)	**92.6**	**86.7**	**90**

Table 2. Comparative results obtained by 1) the baseline system (speaker-Independent); 2) the system using gender-dependent acoustic-phonetic modelling obtained after our Automatic male/female classification. (W.A.: Word Accuracy)

This gender-dependent acoustic-phonetic modelling, based on our new male/female classification, provides very encouraging results. An improvement of 14% in word accuracy rate is observed on the testing set. The reduction of word error rate is observed on substitution, insertion and deletion errors, as seen in table 3.

	Baseline System	Gender-dependent modelling
word error (%)	11.6	10.0
Insertion (%)	1.2	1.0
Deletion (%)	3.8	3.1
Substitution (%)	6.5	5.9

Table 3. Details of word error rate produced by gender-dependent acoustic phonetic modelling in comparison with those obtained with our baseline system.

6. CONCLUSION AND PERSPECTIVES

We have proposed a new automatic male/female classification based on the location of the first and second formants in the frequency domain. This classification is used to split up automatically the training corpora into male and female speakers. Each set of speakers provides the corresponding gender-dependent acoustic models used in the recognition which was implemented on the INRS speech-recognizer system. With these gender-dependent acoustic phonetic models, we reduce by 14% the word error rate.

The next step of this work would be parameter normalization. This procedure will allow us to use only one set of acoustic-phonetic models.

REFERENCES

[1] G. Peterson & H. Barney, "Control methods used in a study of vowels", Journal of the Acoustical Society of America, 24, 175-184, 1952.

[2] R.W. Schafer, L.R. Rabiner, "System for automatic formant analysis of voiced speech", Journal of the Acoustical Society of America, 47, 634-648, 1970.

[3] P. Kenny, R. Hollan, G. Boulianne, H. Garudadri, Y.M. Cheng, M. Lennig, D. O'Shaughnessy, "Experiments in Continuous Speech Recognition witha 60,000 Words vocabulary", *International Conference on Spoken Language Processing*, pp. 225-228, Banff, 1992.

[4] R. Vergin, "Certains Aspects de la perception et l'analyse de la parole pour la reconnaissance automatique", Ph.D. Thesis, INRS-Télécommunications, March 1996.