Automatic Gender Recognition

DAT TRAN and DHARMENDRA SHARMA
School of Information Sciences and Engineering
University of Canberra
Canberra, ACT 2601
AUSTRALIA
{Dat.Tran, Dharmendra.Sharma}@ise.canberra.edu.au

Abstract: This paper presents an automatic gender recognition technique based on speaker's voice. Utterances spoken by same-gender speakers were used to train a text-independent hidden Markov gender model. Female and male models are continuous hidden Markov models. Experiments on the TI-46 database containing 46 words spoken by 8 female and 8 male speakers showed an acceptable recognition rate.

1 Introduction

Most current automatic speech recognition systems are highly speaker dependent. Parametric representations and their probability distributions suitable for a certain speaker may not be suitable for other speakers. It can be said that speaker attributed variability is undesirable in speaker-independent speech recognition systems. The speaker's gender is one of the influential sources of this variability [1]. It is known that the speech recognition performance for female speakers is almost worse than that for male speakers [13]. To improve the performance of speaker-independent speech recognition systems, separate female and male speech models should be used. For example, the performance of the SPHINX-II ASR system improved from adding gender-dependent parameters [6]. Therefore automatic gender recognition is investigated in this paper.

There were several methods proposed for automatic gender recognition, for example, the use of the average value of the fundamental frequency F0, the use of the location in the frequency domain of the first 3 formants for vowels, and minimum vector quantisation distortion method [10, 13].

In this paper, we present an automatic gender recognition technique based on speaker's voice. Utterances spoken by same-gender speakers were used to train a text-independent hidden Markov gender model. Female and male models are left-to-right continuous 5-state 2-mixture hidden Markov models. Experiments on the TI-46 database containing 46 words spoken by 8 female and 8 male speakers showed an acceptable recognition rate.

2 Hidden Markov Models

The underlying assumption of the HMM is that the speech signal can be well characterised as a parametric random process, and that the parameters of the stochastic process can be estimated in a precise, well-defined manner. The HMM method provides a reliable way of recognizing speech for a wide range of applications [8, 11]. The hidden Markov model is a doubly stochastic process with an underlying Markov process which is not directly observable (hidden) but which can be observed through another set of stochastic processes that produce observable events in each of the states [12].

2.1 Parameters of HMMs

Let $O = (o_1\, o_2 \ldots o_T)$ be the observation sequence, $S = (s_1\, s_2 \ldots s_T)$ the unobservable state sequence, $X = (\mathbf{x}_1\, \mathbf{x}_2 \ldots \mathbf{x}_T)$ the continuous vector sequence, $V = \{v_1, v_2, \ldots, v_K\}$ the discrete

symbol set, and N the number of states. A compact notation $\lambda = \{\pi, A, B\}$ is proposed to indicate the complete parameter set of the HMM [12], where

- $\pi = \{\pi_i\}, \quad \pi_i = P(s_1 = i|\lambda), \quad 1 \leq i \leq N$: the initial state distribution;
- $A = \{a_{ij}\}, \quad a_{ij} = P(s_{t+1} = j|s_t = i, \lambda)$, $1 \leq i, j \leq N$, and $1 \leq t \leq T-1$: the state transition probability distribution, denoting the transition probability from state i at time t to state j at time $t+1$; and
- $B = \{b_j(o_t)\}, \quad b_j(o_t) = P(o_t|s_t = j, \lambda)$, $1 \leq j \leq N$, and $1 \leq t \leq T$: the observation probability distribution, denoting the probability of generating an observation o_t in state j at time t with probability $b_j(o_t)$.

2.2 Left-to-Right Continuous HMMs

Left-to-right HMMs: As time increases, the state index increases or stays the same. The state sequence must begin in state 1 and end in state N, i.e. $\pi_i = 0$ if $i \neq 1$ and $\pi_i = 1$ if $i = 1$. The state-transition coefficients satisfy the following fundamental properties

$$a_{ij} = 0 \qquad j < i, \qquad 0 \leq a_{ij} \leq 1$$
$$\text{and} \qquad \sum_{j=1}^{N} a_{ij} = 1 \tag{1}$$

Continuous HMMs (CHMMs): The observations $o_t \in O$ are vectors $\mathbf{x}_t \in X$ and the parametric representation of the observation probabilities is a mixture of Gaussian distributions

$$B = \{b_j(\mathbf{x}_t)\}, \quad 1 \leq j \leq N, \quad 1 \leq t \leq T$$
$$b_j(\mathbf{x}_t) = P(\mathbf{x}_t|s_t = j, \lambda) = \sum_{k=1}^{K} w_{jk} N(\mathbf{x}_t, \mu_{jk}, \mathbf{\Sigma}_{jk})$$
$$\int_{\mathbf{X}} b_j(\mathbf{x}_t) d\mathbf{x}_t = 1 \tag{2}$$

where w_{jk} is the kth mixture weight in state j satisfying $\sum_{k=1}^{K} w_{jk} = 1$ and $N(\mathbf{x}_t, \mu_{jk}, \mathbf{\Sigma}_{jk})$ is the kth Gaussian component density in state j with mean vector μ_{jk} and covariance matrix $\mathbf{\Sigma}_{jk}$.

2.3 Training Left-to-Right Continuous Hidden Markov Gender Models

This problem determines the optimal model parameters λ of the HMM according to given optimisation criterion. The Baum-Welch algorithm yields an iterative procedure to reestimate the model parameters λ using the maximum likelihood criterion [12]. In the Baum-Welch algorithm, the unobservable data are the state sequence S and the observable data are the continuous vector sequence X.

$$\overline{\pi}_j = \gamma_1(i) \tag{3}$$

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{4}$$

$$\overline{w}_{jk} = \frac{\sum_{t=1}^{T} \eta_t(j,k)}{\sum_{t=1}^{T} \sum_{k=1}^{K} \eta_t(j,k)} \tag{5}$$

$$\overline{\mu}_{jk} = \frac{\sum_{t=1}^{T} \eta_t(j,k) \mathbf{x}_t}{\sum_{t=1}^{T} \eta_t(j,k)}, \tag{6}$$

$$\overline{\mathbf{\Sigma}}_{jk} = \frac{\sum_{t=1}^{T} \eta_t(j,k) (\mathbf{x}_t - \overline{\mu}_{jk})(\mathbf{x}_t - \overline{\mu}_{jk})'}{\sum_{t=1}^{T} \eta_t(j,k)} \tag{7}$$

where

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \tag{8}$$

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(x_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(x_{t+1}) \beta_{t+1}(j)} \tag{9}$$

$$\eta_t(j,k) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)} \times \frac{w_{jk}N(\mathbf{x}_t,\mu_{jk},\mathbf{\Sigma}_{jk})}{\sum_{k=1}^{K} w_{jk}N(\mathbf{x}_t,\mu_{jk},\mathbf{\Sigma}_{jk})} \tag{10}$$

the *forward* variable $\alpha_t(i)$ is defined as

$$\alpha_1(i) = \pi_i b_i(\mathbf{x}_1), \quad 1 \le i \le N \tag{11}$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N}\alpha_t(i)a_{ij}\right] b_j(\mathbf{x}_{t+1}) \tag{12}$$

and the *backward* variable $\beta_t(i)$ is defined as

$$\beta_T(i) = 1, \quad 1 \le i \le N \tag{13}$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}b_j(\mathbf{x}_{t+1})\beta_{t+1}(j) \tag{14}$$

Note that for practical implementation, a scaling procedure [12] is required to avoid number underflow on computers with ordinary floating-point number representations.

2.4 Gender Regconition

Assuming that two gender models, female and male, were trained. For each input utterance, its speech signal is converted into a continuous vector sequence X. The probabilities $P(X|\lambda_i)$, $i = 1, 2$ are calculated and the recognised gender is the gender whose probability is highest.

The probability $P(X|\lambda)$ can be computed following both the forward and backward variables as follows

$$P(X|\lambda) = \max_{1\le i\le N} \alpha_t(i)\beta_t(i) \tag{15}$$

This likelihood is computed using the Viterbi algorithm to find the single best state sequence by the maximum operation.

3 Experimental Results

The TI46 corpus was designed and collected at Texas Instruments (TI). The speech was produced by 16 speakers, 8 females and 8 males, labelled f1-f8 and m1-m8 respectively, consisting of two vocabularies—TI-20 and TI-alphabet. The TI-20 vocabulary contains the ten digits from 0 to 9 and ten command words: *enter, erase, go, help, no, rubout, repeat, stop, start*, and *yes*. The TI-alphabet vocabulary contains the names of the 26 letters of the alphabet from *a* to *z*. For each vocabulary item, each speaker produced 10 tokens in a single training session and another two tokens in each of 8 testing sessions. By comparison, the TI-alphabet is a much more difficult vocabulary since it contains several confusable subsets of letters, such as the E-set $\{b, c, d, e, g, p, t, v, z\}$ and the A-set $\{a, j, k\}$. The TI-20 vocabulary is a good choice because it has been used for other tests and therefore can serve as a standard benchmark of the performance. The corpus was sampled at 12500 samples per second and 12 bits per sample.

Speech processing was performed using HTK [14], a toolkit for building hidden Markov models (HMMs). The data were processed in 32 ms frames at a frame rate of 10 ms. Frames were Hamming windowed and preemphasised with $m_p = 0.97$. The basic feature set consisted of 12th-order mel-frequency cepstral coefficients (MFCCs) and the normalised short-time energy, augmented by the corresponding delta MFCCs to form a final set of feature vector with a dimension of 26 for individual frames.

In the training phase, 3680 training tokens (46 words x 8 same gender speakers x 1 training session x 10 repetitions) were used to train a left-to-right continuous hidden Markov gender model (CHMM) for that gender by using the HTK. The speech signals for training CHMMs were converted into feature vector sequences using the linear predictive coding (LPC) analysis. These vector sequences were directly used as observation sequences to train 5-state 2-mixture left-to-right CHMMs.

In the recognition phase, 11776 test utterances (46 words x 16 speakers x 8 test sessions x 2 repetitions) were used to test the two female and male models. We also used the HTK to compute the Viterbi scores and output the gender recognition rate. The gender recognition rates are shown in Table 1. The highest error rate is 2.6 % for the

Table 1: Gender recognition error rates (%) performed on 16 speakers, 11776 test utterances, using left-to-right continuous 5-state 2-mixture hidden Markov gender models

Speakers	Gender Recognition Error Rates (%)
f1	0.2
f2	0.8
f3	0.4
f4	0.6
f5	0.0
f6	0.0
f7	0.0
f8	0.4
m1	0.2
m2	0.7
m3	0.0
m4	1.1
m5	1.3
m6	0.0
m7	2.6
m8	0.2
Female	**2.4**
Male	**6.1**
All	**8.5**

male speaker m7 and the lowest error rate is 0.0% for the female speakers f5, f6, and f7 and the male speakers m3 and m6.

4 Conclusion

We have presented an automatic gender recognition technique based on speaker's voice. Female and male models are left-to-right 5-state 2-mixture continuous hidden Markov models. Experiments on the TI-46 database containing 46 words spoken by 8 female and 8 male speakers showed an acceptable recognition rate of 8.5 %. Large speech and speaker databases and continuous hidden Markov models with different states and mixtures will be considered in our further investigation.

References

[1] W. H. Abdulla and N. K. Kasabov, "Improving speech recognition performance through gender separation", Artificial Neural Networks and Expert Systems International Conference (ANNES), pp 218-222, Dunedin, New Zealand, 2001.

[2] S. Anderson et al. "Recognition of Elderly Speech and Voice-Driven Document Retrieval", in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Phoenix, Arizona, 1999.

[3] D.G. Childers and K. Wu, "Gender recognition from speech". Journal of the Acoustical Society of America, vol. 4, no. 90, pp. 1841-1856, 1991.

[4] R. O. Duda and P. E. Hart, *Pattern classification and scene analysis*, John Wiley & Sons, New York, 1973.

[5] S. Furui, "An Overview of Speaker Recognition Technology", chapter 2 in *Automatic Speech and Speaker Recognition, Advanced Topics*, edited by Chin-Hui Lee, Frank K. Soong, and Kuldip K. Paliwal, Kluwer Academic Publishers, USA, pp. 31-56, 1996.

[6] X.D. Huang et al. "Improved acoustic modeling for the SPHINX speech recognition system" in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 345-348, 1991, Toronto, Canada.

[7] X. D. Huang, Y. Ariki and M. A. Jack, *Hidden Markov Models For Speech Recognition*, Edinburgh University Press, 1990.

[8] B.-H. Juang, "The Past, Present, and Future of Speech Processing", *IEEE Signal Processing Magazine*, vol. 15, no. 3, pp. 24-48, 1998.

[9] S. E. Parris and M. J. Carey, "Language Independent Gender Identification", in Proceedings of the International Conference on

Acoustics, Speech, and Signal Processing (ICASSP), vol. 2, pp. 685-688, 1996.

[10] G. Peterson and H. Barney, "Control Methods used in a Study of Vowels", Journal of the Acoustical Society of America, vol. 24, pp. 174-184, 1952.

[11] L. R. Rabiner, B. H. Juang and C. H. Lee, "An Overview of Automatic Speech Recognition", chapter 1 in *Automatic Speech and Speaker Recognition, Advanced Topics*, edited by Chin-Hui Lee, Frank K. Soong, and Kuldip K. Paliwal, Kluwer Academic Publishers, USA, pp. 1-30, 1996.

[12] L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition*, Prentice Hall PTR, USA, 1993.

[13] R. Vergin, A. Farhat and D. O'Shaughnessy, "Robust Gender-Dependent Acoustic-Phonetic Modelling in Continuous Speech Recognition Based on a New Automatic Male-Female Classification", in Proceedings of International Conferences on Spoken Language Processing (ICSLP), vol. 2, pp. 1081-1084, 1996.

[14] Woodland P. C., Hain T., Johnson S.E., Niesler T.R., Tuerk A., Whittaker E.W.D. and Young S.J. "The 1997 HTK Broadcast News Transcription System", *Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 41-48, Lansdowne.