Australian Accent-Based Speaker Classification

Phuoc Nguyen, Dat Tran, Xu Huang, and Dharmendra Sharma
Faculty of Information Sciences and Engineering
University of Canberra
ACT 2601, Australia
{phuoc.nguyen, dat.tran, xu.huang, dharmendra.sharma}@canberra.edu.au

***Abstract*—This paper presents a new speaker classification scheme based on Australian accents which are broad, general and cultivated. Speakers are classified in to speaker groups according to their accents, ages and genders. Mel-frequency cepstral coefficients extracted after speech processing were used to build Gaussian speaker group mixture models. Fusion of speaker group classifiers is then performed. Experiments showed high performance for the proposed method.**

Keywords: accent recognition, Australian accent, Gaussian mixture model, fusion model.

I. INTRODUCTION

According to linguists, three main varieties of spoken English in Australia are Broad (spoken by 34% of the population), General (55%) and Cultivated (11%) [1]. They are part of a continuum, reflecting variations in accent. Although some men use the pronunciation, the majority of Australians that speak with the accent are women.

Broad Australian English is usually spoken by men, probably because this accent is associated with Australian masculinity. It is used to identify Australian characters in non-Australian media programs and is familiar to English speakers. The majority of Australians speak with the General Australian accent. Cultivated Australian English has some similarities to British Received Pronunciation, and is often mistaken for it. In the past, the cultivated accent had the kind of cultural credibility that the broad accent has today. For example, until 30 years ago newsreaders on the government funded ABC had to speak with the cultivated accent [2].

Although the accent is only spoken by a minority of the population, it has a great deal of cultural credibility. It is disproportionately used in advertisements and by newsreaders. Current research on Australian accent and dialect is focusing on linguistic approach to dialect of phonetic study [3][4], classification of native and non-native Australian [5], or to improve Australian automatic speech recognition performance [6] [7]. However, there is no research on automatic speaker classification based on the three Australian accents of Broad, General, and Cultivated. This speech information as well as gender, age, and emotion are called voice signatures, which are useful for speech mining applications or for the design of a natural spoken-dialog system [8].

Current 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 [1]. For example, the speech recognition performance for female speakers is almost worse than that for male speakers [9]. 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 [10].

In this paper, we propose an automatic speaker classification scheme based on the three Australian accents which are Broad, General, and Cultivated. We also consider the influence of speaker age and gender on the accent classification performance. We use a very large Australian speech corpus ANDOSL [11] in our experiments. There are 18 speaker groups to be classified based on three accents (broad, general, and cultivated), three age ranges (young, middle, and elder), and two genders (female and male). Each group contains 6 speakers. Each speaker speaks 200 utterances and we use 20 of those for training Gaussian mixture models and the remaining 180 utterances for classification. The selection of 20 training utterances is random and is repeated 10 times. The speaker classification method is text-independent, i.e. the text used to train and test the system is completely unconstrained. The Gaussian parameters which are mean vector, covariance matrix and mixture weight are trained in an unsupervised classification using the expectation maximisation (EM) algorithm [12]. Experiments have shown that as long as the training samples cover a sufficient variety of the speaker's speech sound, GMMs are effective models capable of achieving high identification rates for short utterance lengths from unconstrained speech [13].

The rest of the paper is organised as follows. The GMM method and its use for speaker classification are summarised in Section 2. Section 3 presents our experimental results and Section 4 concludes our work.

II. GAUSSIAN MIXTURE MODEL

Let $X = \{x_1, x_2, ..., x_T\}$ be a set of T vectors, each of which is a d-dimensional feature vector extracted by digital speech signal processing. Since the distribution of these vectors is unknown, it is approximately modelled by a mixture of Gaussian densities, which is a weighted sum of K component densities, given by the equation

$$p(x_t \mid \lambda) = \sum_{i=1}^{K} w_i N(x_t, \mu_i, \Sigma_i) \quad (1)$$

978-0-7695-3923-2/10 $26.00 © 2010 IEEE
DOI 10.1109/WKDD.2010.80

where λ denotes a prototype consisting of a set of model parameters $\lambda = \{w_i, \mu_i, \Sigma_i\}$, w_i, i = 1,..., K, are the mixture weights and $N(x_t, \mu_i, \Sigma_i)$, i = 1,..., K, are the d-variate Gaussian component densities with mean vectors μ_i and covariance matrices Σ_i

$$N(x_t, \mu_i, \Sigma_i) = \frac{\exp\left\{-\frac{1}{2}(x_t - \mu_i)'\Sigma_i^{-1}(x_t - \mu_i)\right\}}{(2\pi)^{d/2}\,|\Sigma_i|^{1/2}} \quad (2)$$

In training the GMM, these parameters are estimated such that in some sense, they best match the distribution of the training vectors. The most widely used training method is the maximum likelihood (ML) estimation. For a sequence of training vectors X, the likelihood of the GMM is

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(x_t \mid \lambda) \quad (3)$$

The aim of ML estimation is to find a new parameter model $\overline{\lambda}$ such that $p(X \mid \overline{\lambda}) \geq p(X \mid \lambda)$. Since the expression in (3) is a nonlinear function of parameters in λ, its maximisation is not possible. However, parameters can be obtained iteratively using the expectation-maximisation (EM) algorithm [12]. An auxiliary function Q is used

$$Q(\lambda, \overline{\lambda}) = \sum_{i=1}^{T} p(i \mid x_t, \lambda) \log[\overline{w}_i N(x_t, \overline{\mu}_i, \overline{\Sigma}_i)] \quad (4)$$

where $p(i \mid x_t, \lambda)$ is the *a posteriori* probability for acoustic class i, i = 1,...,c and satisfies

$$p(i \mid x_t, \lambda) = \frac{w_i N(x_t, \mu_i, \Sigma_i)}{\sum_{k=1}^{c} w_k N(x_t, \mu_k, \Sigma_k)} \quad (5)$$

The basis of the EM algorithm is that if $Q(\lambda, \overline{\lambda}) \geq Q(\lambda, \lambda)$ then $p(X \mid \overline{\lambda}) \geq p(X \mid \lambda)$ [10]. The following reestimation formulas are found

$$\overline{w}_i = \frac{1}{T}\sum_{t=1}^{T} p(i \mid x_t, \lambda) \quad (6)$$

$$\overline{\mu}_i = \frac{\sum_{t=1}^{T} p(i \mid x_t, \lambda) x_t}{\sum_{t=1}^{T} p(i \mid x_t, \lambda)} \quad (7)$$

$$\overline{\Sigma}_i = \frac{\sum_{t=1}^{T} p(i \mid x_t, \lambda)(x_t - \overline{\mu}_i)(x_t - \overline{\mu}_i)'}{\sum_{t=1}^{T} p(i \mid x_t, \lambda)} \quad (8)$$

The algorithms for training and classification are described as follows.

Speaker Group Training Algorithm:

Step 1: Generate $p(i \mid x_t, \lambda)$ at random satisfying (5)
Step 2: Compute the mixture weight, the mean vector, and the covariance matrix following (6), (7) and (8)
Step 3: Update $p(i \mid x_t, \lambda)$ according to (5) and compute the function Q using (4)
Step 4: Stop if the increase in the value of the function Q at the current iteration relative to the value of the Q function at the previous iteration is below a chosen threshold, otherwise go to step 2.

Speaker Group Classification Algorithm:

Let λ_k, k = 1,…, N, denote accent models of N speaker groups. Given a feature vector sequence X, a classifier is designed to classify X into N speaker groups by using N discriminant functions $g_k(X)$, computing the similarities between the unknown X and each speaker group model λ_k and selecting the model λ_{k^*} if

$$k^* = \arg\max_{1 \leq k \leq N} g_k(X) \quad (9)$$

In the minimum-error-rate classifier, the discriminant function is the *a posteriori* probability. Using the Bayes rule and assuming equally likely speakers, the discriminant function in (10) is equivalent to the following

$$g_k(X) = p(X \mid \lambda_k) \quad (10)$$

Finally, using the log-likelihood, the decision rule used for speaker identification is

Select speaker group k if*

$$k^* = \arg\max_{1 \leq k \leq N} \sum_{t=1}^{T} \log p(x_t \mid \lambda_k) \quad (11)$$

where $p(x_t \mid \lambda_k)$ is given in (1).

III. Experimental Results

A. ANDOSL Database

The Australian National Database of Spoken Language (ANDOSL) corpus [11] comprises carefully balanced material for Australian speakers, both Australian-born and overseas-born migrants. The aim was to represent as many significant speaker groups within the Australian population as possible. Current holdings are divided into those from native speakers of Australian English (born and fully educated in Australia) and those from non-native speakers of Australian English (first generation migrants having a non-English native language). A subset used for speaker verification experiments in this paper consists of 108 native

speakers. There are 36 speakers of General Australian English, 36 speakers of Broad Australian English and 36 speakers of Cultivated Australian English in this subset. Each of the three groups comprises 6 speakers of each gender in each of three age ranges (18-30, 31-45 and 46+). So there are total of 18 groups of 6 speakers labeled *ijk*, where *i* denotes *f* (female) or *m* (male), *j* denotes *y* (young) or *m* (medium) or *e* (elder), and *k* denotes *g* (general) or *b* (broad) or *c* (cultivated). For example, the group *fyg* contains 6 female young general Australian English speakers. Each speaker contributed in a single session, 200 phonetically rich sentences. All waveforms were sampled at 20 kHz and 16 bits per sample.

B. *Speech Processing*

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 pre-emphasised 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.

GMMs are initialized as follows. Mixture weights, mean vectors, and covariance matrices were initialized with essentially random choices. Covariance matrices are diagonal, i.e. $[\sigma_k]_{ii} = \sigma_k^2$ and $[\sigma_k]_{ij} = 0$ if $i \neq j$, where σ_k^2, $1 \leq k \leq K$ are variances. A variance limiting constraint was applied to all GMMs using diagonal covariance matrices [13]. This constraint places a minimum variance value $\sigma_{\min}^2 = 10^{-2}$ on elements of all variance vectors in the GMM in our experiments.

C. *Accent Classification Results versus Number of Gaussian Components*

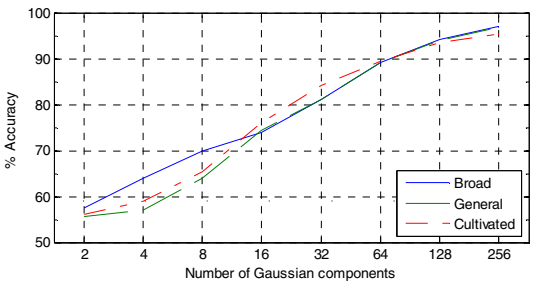

Figure 1. Accent classification for Broad, General and Cultivated groups

Figure 1 presents the classification rate averaged on 10 experiments where the 20 training utterances were randomly selected. Overall the classification rates are higher when the number of Gaussian components increases. The Cultivated accent gets better results for 15 Gaussians or higher and achieves the highest classification rate of 96% for 256 Gaussians.

The standard deviation (STDEV) was measured to consider how widely values are dispersed from the average value. Low STDEV indicates that the values tend to be very close to the mean and the accuracies are consistent when repeating experiments. Table I shows the STDEV of the accent classification rates for the 10 experiments. The results are consistent for 256 Gaussians.

TABLE I. STANDARD DEVIATION (%) OF ACCENT CLASSIFICATION FROM 10 EXPERIMENTS

	2	4	8	16	32	64	128	256
Broad	1.61	2.39	2.00	1.55	1.24	0.82	0.52	0.34
General	2.91	3.65	4.27	2.74	2.13	1.34	0.92	0.62
Cultivated	2.62	2.09	3.76	2.21	1.51	1.02	0.65	0.61

D. *Accent Classification Results versus Age and Gender*

We consider the influence of age and gender on the accent classification. We divide 108 speakers in to 18 speaker groups based on the three accents Broad, General, and Cultivated, three ages Young, Middle, and Elderly, and two genders Male and Female. Each group contains 6 speakers. The number of Gaussians was set to 256.

Figure 2 shows the accent classification versus age. While the classification rates of Broad and General slightly increase from 94% and 96% at Young age to 98% and 97%, respectively for Middle age and Elderly age, the accuracy of Cultivated is dropped down from 98% for Young to 94% for Middle and to 89% for Elderly. The best results were found for the Middle group. These show that the accent is best recognized for middle speakers and the Cultivated accent is hard to recognize for elderly speakers.

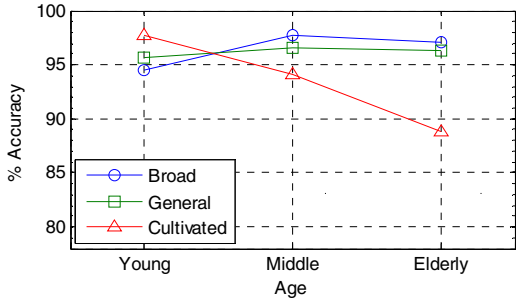

Figure 2. Accent classification versus age

Similar to the previous classification results, we also considered the robustness of this classification by calculating the STDEV values on 10 experiments and listed them on Table II below. The values are very low, ranging from 0.39% to 1.41%, which guarantee the robustness.

TABLE II. STANDARD DEVIATION (%) OF ACCENT CLASSIFICATION VERSUS AGE AVERAGED ON 10 EXPERIMENTS

	DA			DAM			DAF		
	Y	M	E	Y	M	E	Y	M	E
Broad	0.79	0.60	0.51	1.52	0.85	0.98	0.50	0.67	0.16
Cultivated	0.48	1.41	0.39	0.68	0.14	0.40	0.81	2.73	0.78
General	1.16	0.36	1.02	0.99	0.68	0.35	2.26	0.48	1.91

Figures 3 and 4 show the accent classification rates versus age performed on male and female speakers, respectively. The Cultivated accent is recognizable on male speakers only.

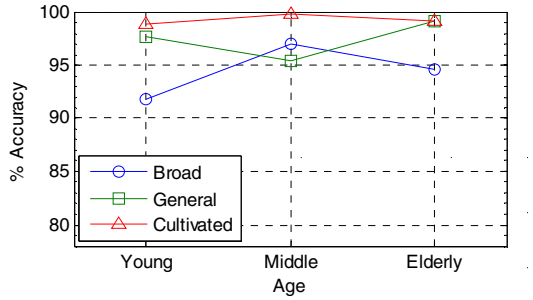

Figure 3. Accent classification versus age performed on male speakers

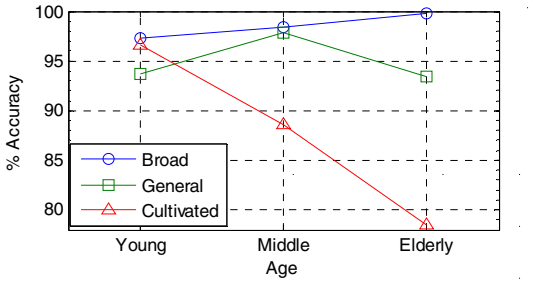

Figure 4. Accent classification versus age performed on female speakers

IV. CONCLUSION

We have presented speaker classification based on the three Australian accents which are Broad, General and Cultivated using the ANDOSL database consisting of 108 speakers, each speaks 200 long utterances. We considered the classification rates versus the number of Gaussian components, age, and gender. We extracted MFCC features for speech and used those features to train Gaussian speaker models. Most classification rates were high, ranging from 90% to 99%. The Cultivated accent is hard to recognise for elderly female speakers.

REFERENCES

[1] Mitchell and Delbridge 1965, The Pronunciation of English in Australia, pp. 11–19

[2] http://www.convictcreations.com/research/languageidentity.html

[3] J. Harrington, F. Cox, and Z. Evans, "An acoustic phonetic study of broad, general, and cultivated Australian English vowels," Australian Journal of Linguistics, vol. 17, no. 2, pp. 155-184, 1997.

[4] K. Berkling, M. Zissman, J. Vonwiller, and C. Cleirigh, "Improving accent identification through knowledge of English syllable structure," in ICSLP-1998, 1998, pp. 89-92.

[5] K. Kumpf and R. W. King, "Automatic accent classification of foreign accented Austrialian english speech," in Fourth Internalional Conference on Spoken Language Processing, 1996, pp. 1740-1743

[6] A. S. Kollengode, H. Ahmad, B. Adam, and B. Serge, "Performance of Speaker-independent Speech Recognisers for Automatic Recognition of Australian English," in Proceedings of the 11th Australian International Conference on Speech Science & Technology, Auckland, 2006, pp. 494-499

[7] B.R. Wildermoth and K.K. Paliwal, "Gmm based speaker recognition on," in Micro.Elec.Eng. Research Conf., 2003

[8] I. Shafran, M. Riley, and M. Mohri. 2003. Voice signatures. In Proc. IEEE Automatic Speech Recognition and UnderstandingWorkshop.

[9] 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.

[10] 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.

[11] J. B. Millar, J. P. Vonwiller, J. M. Harrington, and P. J. Dermody, "The Australian National Database of Spoken Language", in Proc. Int. Conf. Acoust., Speech, Signal Processing (ICASSP94), 1, pp. 97-100, 1994.

[12] R. Hathaway, "Another interpretation of the EM algorithm for mixture distribution", Journal of Statistics & Probability letters, vol. 4, pp. 53-56, 1986.

[13] D. A. Reynolds, "Robust text-independent speaker identification using Gaussian mixture models", IEEE Transactions on Speech and Audio Processing, 3:1, pp. , January 1995.

[14] P. C. Woodland, "Broadcast news transcription using HTK", in Proceedings of International Conference on Acoustics, Speech & Signal Processing (ICASSP97), pp. , USA, 1997.

[15] 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.

[16] Campbell, J. P., "Speaker Recognition: A Tutorial", Special issue on Automated biometric Systems, Proceedings of IEEE, 85:9, pp. 1436-1462, September, 1997.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM algorithm", Journal of the Royal Statistical Society, Ser. B, 39: pp. 1-38, 1977.

[18] R. O. Duda and P.E. Hart, "Pattern classification and scene analysis", John Wiley & Sons, 1973.

[19] Furui, S., "Recent advances in speaker recognition", Patter Recognition Letters, 18, pp. 859-872, 1997.

[20] S. Furui, "An overview of speaker recognition technology", ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 1-9, 1994.

[21] J. Harrington, F. Cox & Z. Evans, "An acoustic study of broad, general and cultivated Australian English vowels", Australian Journal of Linguistics, 17:2, pp. 155-184, 1996.

[22] B. H. Juang, "The Past, Present, and Future of Speech Processing", IEEE Signal Processing Magazine, 15:3, pp. 24-48, May, 1998.

[23] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs", Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, San Francisco, pp. II-157-160, 1960.

[24] D. Tran and M.Wagner, "Fuzzy Gaussian Mixture Models for Speaker Recognition", Australian Journal of Intelligent Information Processing Systems (AJIIPS), vol. 5, no. 4, pp. 293-300, 1998.

[25] D. Tran and M.Wagner, "Fuzzy Approach to Gaussian Mixture Models and Generalised Gaussian Mixture Models", Proc. Computation Intelligence Methods and Applications (CIMA99), USA, 1999.