# Automatic Classification of Speaker Characteristics

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*—**An automatic voice-based classification system of speaker characteristics including age, gender and accent is presented in this paper. Speakers are grouped according to their characteristics and their speech features are then extracted to train speaker group models using different classification techniques. Finally fusion of classification results for those speaker groups is performed to obtain results for each speaker characteristic. The ANDOSL Australian speech database consisting of 108 speakers and 21600 long utterances was used for system evaluation. Experiments showed high performance for the proposed classification of speaker characteristics.**

***Keywords-component; speaker characteristics; speech processing; vector quantization; Gaussian mixture model; support vector machine.***

## I. INTRODUCTION

Speaker characteristics can be divided in to relatively stable characteristics and transient characteristics. Stable speaker characteristics comprise physiological and anatomical factors such as gender and age. Transient speaker characteristics comprise stress and emotional state. Stable speaker characteristics are easier to recognise [1]. The most important stable speaker characteristics will be mentioned below.

Classifying speaker characteristics is an important task in Dialog Systems, Speech Synthesis, Forensics, Language Learning, Assessment Systems, and Speaker Recognition Systems [2]. In *Human-Computer Interaction* applications, the interaction between users and computers taking place at the speech-driven user interface. For example, Spoken Dialogs Systems provide services in domains of finance, travel, scheduling, tutoring, or weather. The systems need to gather automatically information from the user in order to provide timely and relevant services. Most telephone-based services today use spoken dialog systems to either route calls to the appropriate agent or even handle the complete service by an automatic system. Another example of *Human-Computer Interaction* application is Computer-aided Learning and Assessment systems. The systems provide interactive recording and playback of user's input speech, feedback regarding acoustic speech features, recognizing the input, and interpreting interaction to act as a conversation partner. Besides customizing to the native language of the language learner, learning systems may have to be tailored towards particular accents, for example the E-Language Learning System program between the U.S. Department of Education and the Chinese Ministry of Education. In *Human-Centered* applications, the computers stay in the background attempting to anticipate and serve people's needs. One example is Smart Room Environments in which computers watch and interpret people's actions and interactions in order to support communication goals. Another example is Speech Translation system whose task is to recognize incoming speech from the source language and translate the text of the recognizer output into text of the target language, and then synthesize the translated text to audible speech in the target language. The system needs to generate appropriate synthesized output based on the speaker's gender, age and accent. Beyond that, speaker characteristics need to be assessed in order to adapt system components, particularly the speech recognition front-end to the specific voice characteristics of the speaker and the content of what was spoken. This adaptation process has been proven to dramatically improve the recognition accuracy, which usually carries over favorably to the performance of the overall system. Recent systems rely on speaker adaptive training methods, which first determine the speaker's identity and then apply acoustic model adaptation based on the assumed identity. Some applications rely on broader speaker classes such as gender or age to load pre-trained models [2].

A number of investigations on speaker characteristics have been found in the literature. Elderly speakers were identified in [3] using Gaussian mixture models. In [4] general acoustic and prosodic features were also used to train hidden Markov models to classify speaker's gender, age, dialect, and emotion. Experiments in [5] used four classifiers for separate recognition of age and gender. In [6]-[10], feature analysis was investigated, results showed prosodic features gain better performance over acoustic features while do not require linguistic features. For accent classification, we particularly focus on Australian accent. 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.

According to linguists, three main varieties of spoken English in Australia are Broad (spoken by 34% of the population), General (55%) and Cultivated (11%) [11]. 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

978-1-4244-7057-0/10/$26.00 ©2010 IEEE

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 [12].

Current research on Australian accent and dialect is focusing on linguistic approach to dialect of phonetic study [13][14], classification of native and non-native Australian [15], or to improve Australian automatic speech recognition performance [16]. However, there is no research on automatic speaker classification based on the three Australian accents of Broad, General, and Cultivated. There has not been a classification system that can classify persons based on their gender, age and accent simultaneously.

This paper presents a scheme of voice-based classification of speaker characteristics. In the first stage, speakers are grouped into 18 speaker groups which are combinations of 2 gender groups (female and male), 3 age groups (young, middle and elderly), and 3 accent groups (broad, general and cultivated). Speech processing was performed using open source openSMILE feature extraction [17]. There are 16 low-level descriptors chosen including zero-crossing-rate (ZCR) from the time signal, root mean square (RMS) frame energy, pitch frequency (normalised to 500 Hz), harmonics-to-noise ratio (HNR) by autocorrelation function, and mel-frequency cepstral coefficients (MFCC) 1-12 in full accordance to HTK-based computation [18]. Voice features are extracted as feature vectors and are used to train speaker group models with different techniques which are Vector Quantization (VQ), Gaussian Mixture Model (GMM), and *C*-Support Vector Classifiers (C-SVC). Fusion of classification results from those groups is then performed to obtain results for each gender, age and accent. The ANDOSL Australian speech database consisting of 108 speakers and 21600 long utterances was used for evaluation [19]. Experiments showed high performance for the proposed classification system.

The rest of the paper is organised as follows. VQ, fuzzy VQ, GMM and *C*-SVC methods are summarised in Section 2. Section 3 presents our experimental results and Section 4 concludes our work.

## II. Classifiers for Speaker Characteristics

### A. Vector Quantization

Vector quantization (VQ) is a data reduction method, which is used to convert a feature vector set into a small set of distinct vectors using a clustering technique [20]. The distinct vectors are called codevectors and the set of codevectors that best represents the training vector set is called the codebook. The VQ codebook can be used as a speech or speaker model. Since there is only a finite number of code vectors, the process of choosing the best representation of a given feature vector is equivalent to quantising the vector and leads to a certain level of quantization error. This error decreases as the size of the codebook increases, however the storage required for a large codebook is nontrivial. The key point of VQ modelling is to derive an optimal codebook which is commonly achieved by using the hard *C*-means (HCM) (k-means).

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. Let $U = [u_{it}]$ be a matrix whose elements are memberships of $x_t$ in the ith cluster, $i=1,...,C$, $t=1,...,T$. Hard $C$-partition space for $X$ is the set of matrices $U$ such that

$$u_{it} \in \{0,1\}\ \forall i,t, \quad \sum_{i=1}^{C} u_{it} = 1\ \forall t, \qquad 0 < \sum_{i=1}^{C} u_{it} < T\ \forall i \quad (1)$$

where $u_{it} = u_i(x_t)$ is 1 or 0 according to whether $x_t$ is or is not in the ith cluster, $\sum_{i=1}^{C} u_{it} = 1\ \forall t$ means each $x_t$ is in exactly one of the $C$ clusters, and $0 < \sum_{i=1}^{C} u_{it} < T = 1\ \forall i$ means that no cluster is empty and no cluster is all of $X$ because of $2 \le C < T$

The HCM method is based on minimisation of the sum-of-squared-errors function as follows [20]

$$J_m(U, \lambda; X) = \sum_{i=1}^{C}\sum_{t=1}^{T} u_{it} d_{it}^2 \quad (2)$$

where $U = \{u_{it}\}$ is a hard $C$-partition of $X$, $\lambda$ is a set of prototypes, in the simplest case, it is the set of cluster centers: $\lambda = \{\mu\}, \mu = \{\mu_i\}, i = 1,...,C$ and $d_{it}$ is the distance in the $A$ norm ($A$ is any positive definite matrix) from $x_t$ to $\mu_i$, known as a measure of dissimilarity

$$d_{it}^2 = \| x_t - \mu_i \|_A^2 = (x_t - \mu_i)' A (x_t - \mu_i) \quad (3)$$

Minimizing the hard objective function $J_m(U, \lambda; X)$ in (2) gives

$$u_{it} = \begin{cases} 1 & d_{it} < d_{jt} \quad j = 1,...,C, \ \forall j \ne i \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$\mu_i = \sum_{t=1}^{T} u_{it} x_t \Big/ \sum_{t=1}^{T} u_{it} \quad (5)$$

where ties are broken randomly.

### B. Fuzzy Vector Quantization

Fuzzy Vector Quantization (FVQ) is a fuzzy partitioning of $X$ into $C$ fuzzy subsets or $C$ clusters, $1 < C < T$. The most important requirement is to find a suitable measure of clusters, referred to as a fuzzy clustering criterion. Objective function methods allow the most precise formulation of the fuzzy clustering criterion. The most well known objective function for fuzzy clustering in $X$ is the least-squares functional, that is, the infinite family of fuzzy $C$-means (FCM) functions, generalized from the classical within-groups sum of squared error function [21][22]

$$J_m(U, \lambda; X) = \sum_{i=1}^{C}\sum_{t=1}^{T} u_{it}^m d_{it}^2 \quad (6)$$

where $U = \{u_{it}\}$ is a fuzzy $c$-partition of $X$, each $u_{it}$ represents the degree of vector $x_t$ belonging to the $i$th cluster

and is called the fuzzy membership function. For $1 \le i \le C$ and $1 \le t \le T$, we have

$$0 \le u_{it} \le 1, \quad \sum_{i=1}^{c} u_{it} = 1, \text{ and } 0 < \sum_{t=1}^{T} u_{it} < T \tag{7}$$

$m > 1$ is a weighting exponent on each fuzzy membership $u_{it}$ and is called the degree of fuzziness; other parameters are defined as seen in VQ.

Minimizing the fuzzy objective function $J_m$ in (6) gives

$$u_{it} = \left[ \sum_{k=1}^{c} (d_{it} / d_{kt})^{\frac{2}{m-1}} \right]^{-1} \tag{8}$$

$$\mu_i = \sum_{t=1}^{T} u_{it}^m x_t \Big/ \sum_{t=1}^{T} u_{it}^m \tag{9}$$

### C. Gaussian Mixture Model

Since the distribution of feature vectors in $X$ 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

$$p(x_t \mid \lambda) = \sum_{i=1}^{K} w_i N(x_t, \mu_i, \Sigma_i) \tag{10}$$

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

$$N(x_t, \mu_i, \Sigma_i) = \frac{\exp\left\{-\frac{1}{2}(x_t - \mu_i)' \Sigma_i^{-1} (x_t - \mu_i)\right\}}{(2\pi)^{d/2} \mid \Sigma_i \mid^{1/2}} \tag{11}$$

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

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(x_t \mid \lambda) \tag{12}$$

The aim of ML estimation is to find a new parameter model $\bar{\lambda}$ such that $p(X \mid \bar{\lambda}) \ge p(X \mid \lambda)$. Since the expression in (12) is a nonlinear function of parameters in $\lambda$, its direct maximisation is not possible. However, parameters can be obtained iteratively using the expectation-maximisation (EM) algorithm [23]. An auxiliary function $Q$ is used

$$Q(\lambda, \bar{\lambda}) = \sum_{i=1}^{T} p(i \mid x_t, \lambda) \log[\bar{w}_i N(x_t, \bar{\mu}_i, \bar{\Sigma}_i)] \tag{13}$$

where $p(i \mid x_t, \lambda)$ is the posterior probability for acoustic class $i$, $i = 1, \ldots, c$ and satisfies

$$p(i \mid x_t, \lambda) = \frac{w_i N(x_t, \mu_i, \Sigma_i)}{\sum_{k=1}^{c} w_k N(x_t, \mu_k, \Sigma_k)} \tag{14}$$

The basis of the EM algorithm is that if $Q(\lambda, \bar{\lambda}) \ge Q(\lambda, \lambda)$ then $p(X \mid \bar{\lambda}) \ge p(X \mid \lambda)$ [24][25][26]. The following re-estimation equations are found

$$\bar{w}_i = \frac{1}{T} \sum_{t=1}^{T} p(i \mid x_t, \lambda) \tag{15}$$

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} p(i \mid x_t, \lambda) x_t}{\sum_{t=1}^{T} p(i \mid x_t, \lambda)} \tag{16}$$

$$\bar{\Sigma}_i = \frac{\sum_{t=1}^{T} p(i \mid x_t, \lambda)(x_t - \bar{\mu}_i)(x_t - \bar{\mu}_i)'}{\sum_{t=1}^{T} p(i \mid x_t, \lambda)} \tag{17}$$

### D. Support Vector Machine

#### 1) Binary Case

Label the training data $\{x_i, y_i\}, i = 1, \ldots, l$, $y_i \in \{-1, 1\}$, $x_i \in R^d$. The support vector machine (SVM) using C-Support Vector Classification (C-SVC) algorithm will find the optimal hyperplane [27]

$$f(x) = w^T \Phi(x) + b \tag{18}$$

to separate the training data by solving the following optimization problem:

$$\min \quad \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i \tag{19}$$

subject to

$$y_i \left[ w^T \Phi(x_i) + b \right] \ge 1 - \xi_i \text{ and } \xi_i \ge 0, i = 1, \ldots, l \tag{20}$$

The optimization problem (19) will guarantee to maximize the hyperplane margin while minimize the cost of error. $\xi_i, i = 1, \ldots, l$ are non-negative slack variables introduced to relax the constraints of separable data problem to the constraint (9) of non-separable data problem. For an error to occur the corresponding $\xi_i$ must exceed unity (20), so $\sum_i \xi_i$ is an upper bound on the number of training errors. Hence an extra cost $C \sum_i \xi_i$ for errors is added to the objective function (19) where $C$ is a parameter chosen by the user.

The Lagrangian formulation of the primal problem is:

$$L_P = \frac{1}{2}\|w\|^2 + C \sum_i \xi_i - \sum_i \alpha_i \{y_i (x_i^T w + b) - 1 + \xi_i\} - \sum_i \mu_i \xi_i \tag{21}$$

We will need the Karush-Kuhn-Tucker conditions for the

primal problem to attain the dual problem:

$$L_D = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \Phi(x_i)^T \Phi(x_i) \quad (22)$$

subject to:

$$0 \le \alpha_i \le C \qquad \text{and} \qquad \sum_i \alpha_i y_i = 0 \quad (23)$$

The solution is given by:

$$w = \sum_i^{N_S} \alpha_i y_i x_i \quad (24)$$

where $N_S$ is the number of support vectors.

Notice that data only appear in the training problem (21) and (22) in the form of dot product $\Phi(x_i)^T \Phi(x_i)$ and can be replaced by any kernel $K$ with $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$, $\Phi$ is a mapping to map the data to some other (possibly infinite dimensional) Euclidean space. One example is Radial Basis Function (RBF) kernel $K(x_i, x_j) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2}$

In test phase an SVM is used by computing the sign of

$$f(x) = \sum_i^{N_S} \alpha_i y_i \Phi(s_i)^T \Phi(x) + b = \sum_i^{N_S} \alpha_i y_i K(s_i, x) + b \quad (25)$$

where the $s_i$ are the support vectors.

*2) Multiclass Support Vector Machine*

The binary SVM classifiers can be combined to handle the multiclass case: One-against-all classification uses one binary SVM for each class to separate their members to other classes, while Pairwise classification uses one binary SVM for each pair of classes to separate members of one class from members of the other.

## III. Experimental Results

### A. ANDOSL Database

The Australian National Database of Spoken Language (ANDOSL) corpus [19] 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

In speaker characteristics feature research, prosodic approaches attempt to capture speaker-specific variation in intonation, timing, and loudness [2]. Because such features are supra-segmental (are not properties of single speech segments but extend over syllables and longer regions), they can provide complementary information to systems based on frame-level or phonetic features. One of the most studied features is speech fundamental frequency (or as perceived, pitch), which reflects vocal fold vibration rate and is affected by various physical properties of the speaker's vocal folds, including their size, mass, and stiffness. Distributions of frame-level pitch values have been used in a number of studies. Although they convey useful information about a speaker's distribution of pitch values, such statistics do not capture dynamic information about pitch contours and are thus not viewed as high-level here [1]-[10].

Speech processing was performed using open source openSMILE feature extraction [17]. There are 16 low-level descriptors chosen including ZCR, RMS energy, pitch frequency, HNR, and MFCC 1-12 in full accordance to HTK-based computation. To each of these, the delta coefficients are additionally computed. Next the 12 functionals including mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, and range as well as two linear regression coefficients with their mean square error (MSE) are applied on a chunk basis. Thus, the total feature vector per chunk contains 16 * 2 * 12 = 384 attributes.

### C. Parameter Settings for VQ and FVQ

Because the feature values have different ranges and the Euclidean distance was used, the following normalization of features was applied:

$$x'_{tj} = \frac{x_{tj} - \mu_j}{s_j} \quad (26)$$

where $x_{tj}$ is the $j$-th feature of the $t$-th vector, $\mu_j$ the mean value of all $T$ vectors for feature $j$, and $s_j$ the absolute standard deviation, that is

$$s_j = \frac{1}{T}\sum_{t=1}^{T} |x_{tj} - \mu_j| \quad (27)$$

In order to find good selection of number of clusters and to watch accuracy trend, various number of them are tried to conduct experiments. Result in Figure 1 shows that the highest accent classification rate is found when the number of clusters is 32.

### D. Parameter Settings for GMM

GMM is regarded as one state continuous hidden Markov model, therefore we used HTK toolkit [18] to train and test GMMs. All feature vectors were converted to HTK format. The number of Gaussians was set to 32, which is equal to the number of code vectors in FVQ and VQ.

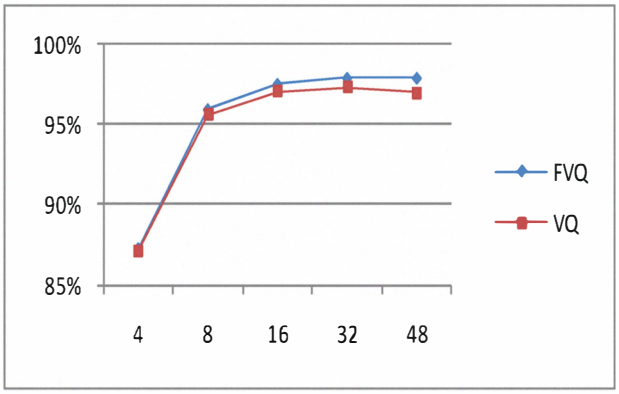


Fig. 1. Classification rates for FVQ (using Fuzzy C-Means) and VQ (using K-means).

### E. *Parameter Settings for SVM*

Experiments were performed using WEKA data mining tool [28][29], C-SVC with RBF kernel were selected. All feature vectors were scale to range [-1, 1] in order to avoid domination of some dimension to general performance of classifiers. We performed several experiments with different values of parameters $C$ and $\gamma$ to search for the best model. The chosen values were $C = 2^1, 2^3, ..., 2^{15}$ and $\gamma = 2^{-15}, 2^{-13}, ..., 2^3$. The 10-fold cross-validation was used with every pair of values of $C$ and $\gamma$. Results are shown in Figure 2 and we can see that the best values are $C = 2^7$ and $\gamma = 2^{-5}$.

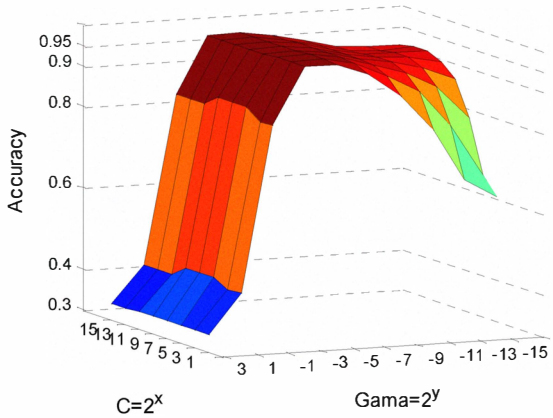


Fig. 2. Accent classification rates versus $C$ and $\gamma$.

### F. *Experimental Results*

We used 10-fold cross validation for evaluation our system. The ANDOSL dataset was divided in to 10 equal subsets where 9 subsets were used to train models and the remaining subset was used for evaluation. We averaged 10 results from 10 times applying the cross validation to obtain the final result for each technique. Table 1 summarizes all of the final results.

Results showed that SVM achieved the best performance for all Gender, Age and Accent classifications. FVQ is better than VQ and GMM. As seen in the Introduction section, Cultivated Australian English has some similarities to British Received Pronunciation, and is often mistaken for it. The results in Table I also show that Cultivated classification achieved the lowest classification rate comparing with the other two accents Broad and General.

TABLE I. CLASSIFICATION RATE (%) FOR ALL CHARACTERISTICS AND TECHNIQUES

| | **Gender** | | **Age** | | | **Accent** | | |
|---|---|---|---|---|---|---|---|---|
| | *Male* | *Female* | *Young* | *Middle* | *Elderly* | *Broad* | *General* | *Cultivated* |
| SVM | 100.0 | 100.0 | 98.6 | 98.7 | 99.0 | 99.0 | 98.7 | 98.3 |
| FVQ | 100.0 | 99.9 | 98.6 | 98.1 | 98.7 | 98.7 | 98.4 | 98.2 |
| VQ | 99.9 | 99.9 | 97.9 | 97.9 | 98.0 | 98.2 | 98.1 | 97.6 |
| GMM | 100.0 | 99.9 | 96.7 | 96.7 | 97.6 | 97.2 | 96.7 | 96.7 |

### G. *More Results for SVM*

TABLE II. CONFUSION MATRICES FOR GENDER CLASSIFICATION

| | **Male** | **Female** |
|---|---|---|
| Male | 10797 | 3 |
| Female | 1 | 10799 |

TABLE III. CONFUSION MATRICES FOR AGE CLASSIFICATION

| | **Young** | **Middle** | **Elderly** |
|---|---|---|---|
| Young | 7097 | 53 | 50 |
| Middle | 50 | 7109 | 41 |
| Elderly | 32 | 37 | 7131 |

TABLE IV. CONFUSION MATRICES FOR ACCENT CLASSIFICATION

| | **Broad** | **General** | **Cultivated** |
|---|---|---|---|
| Broad | 7128 | 27 | 45 |
| General | 25 | 7109 | 66 |
| Cultivated | 67 | 56 | 7077 |

Tables II, III and IV present confusion matrices for each gender, age and accent classification. The total utterances are 21600. Table II shows very good result for gender classification. Table III shows reasonable errors. The number of utterances of young people misclassified as middle age people is higher than that misclassified as elderly people. Similar result is found for misclassified utterances of elderly people. Result for accent classification in Table IV shows the lowest classification rate for Cultivated comparing with the other two accents Broad and General.

## IV. CONCLUSION

We have presented a classification scheme for speaker characteristics including gender, age and accent characteristics. Speakers were divided into subgroups according to their characteristics. Their utterances were used to train speaker group models using vector quantization, Gaussian mixture modelling and support vector machine techniques. From the classification results of those speaker groups, a fusion of the results was implemented to obtain classification results for each characteristic. The proposed classification scheme was

evaluated on the Australian speech database consisting of 108 speakers and 200 utterances for each speaker. Useful speech features were extracted. Most of classification rates were high, ranging from 97.96% to 98.68%. These good results showed that the proposed classification scheme can be used in classification of speaker characteristics.

## REFERENCES


[1] K. Scherer & H. Giles, *Social markers in speech*. Cambridge University Press, Cambridge, 1979.

[2] T. Schultz, "Speaker characteristics," in Speaker Classification I. Springer Berlin / Heidelberg, 2007, pp. 47-74.

[3] N. Minematsu, M. Sekiguchi, and K. Hirose. Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers. In Proc. IEEE Int'l Conference on Acoustic Signal and Speech Processing, pp. 137–140, 2002.

[4] I. Shafran, M. Riley, and M. Mohri. 2003. Voice signatures. In Proc. IEEE Automatic Speech Recognition and Understanding Workshop.

[5] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt,J. Stegmann, C. Müller, R. Huber, B. Andrassy, J.G. Bauer,and B. Littel, "Comparison of Four Approaches to Age and Gender Recognition for Telephone Applications," in ICASSP2007 Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, Hawai'i, USA,2007, vol. 4, pp. 1089 – 1092.

[6] E. Shriberg, "Higher-Level Features in Speaker Recognition," in Speaker Classification I. Springer Berlin / Heidelberg, 2007, pp. 241-259.

[7] S. Schötz, "Acoustic analysis of adult speaker age," in Speaker Classification I. Springer Berlin / Heidelberg, 2007, pp. 88-107.

[8] J. P. Campbell, D. A. Reynolds, and R. B. Dunn, "Fusing high- and low-level features for speaker recognition," in Proceedings of Eurospeech, 2003, pp. 2665–2668.

[9] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt,J.Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous,and V. Aharonson, "The relevance of feature type for the automatic classification of emotional user states: Low Level Descriptors and Functionals," in Proc. Interspeech, Antwerp, 2007, pp. 2253–2256.

[10] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in Proc. Interspeech. Brighton, UK: ISCA, 2009

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

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

[13] 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.

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

[15] 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

[16] 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.

[17] F. Eyben, M. W¨ollmer, B. Schuller (2009): Speech and Music Interpretation by Large-Space Extraction, http://sourceforge.net/projects/openSMILE.

[18] P.C. Woodland, M.J.F. Gales, D. Pye & S.J. Young (1997).Broadcast news transcription using HTK. Proc. ICASSP'97,pp. 719–722, Munich.

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

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

[21] D. Tran, W. Ma, D. Sharma and T. Nguyen, "Fuzzy Vector Quantization for Network Intrusion Detection", IEEE International Conference on Granular Computing, Silicon Valley, 2-4 November 2007, USA.

[22] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press, 1981.

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

[24] X.D. Huang, K. Lee, H. Hon, and M. Hwang, "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.

[25] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEETrans. Speech Audio Processing, vol. 3, no. 1, pp. 72–83, 1995.

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

[27] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," Knowledge Discovery and Data Mining, vol. 2, no. 2, pp.121–167, 1998.

[28] I. H. Witten and E. Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco, 2nd edition, 2005.

[29] C.-C. Chang and C.-J. Lin. LibSVM: a library for sup-port vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm.