GMM Based Speaker Recognition on Readily Available Databases

Brett R. Wildermoth & Kuldip K. Paliwal
School of Microelectronic Engineering, Griffith University, Brisbane, Australia 4111
Email: B.Wildermoth@griffith.edu.au, K.Paliwal@griffith.edu.au

***Abstract*— In this paper we give an overview of the most popular databases used in speaker recognition evaluation and clearly outline the means of training and testing an ASR system using them. The complete ASR system including both feature extraction and classification is explicitly explained in great detail. The performance of a GMM based system for speaker verification and identification is reported using various forms of MFCC features on the TIMIT, YOHO and ANDOSL databases.**

I. Introduction

Automatic Speaker Recognition (ASR) systems are useful for verifying the identity of a person; allow automated control of services by voice, such as banking transactions and also control the flow of confidential information. While retinal scans and fingerprints are considered more reliable means of identification, speech can be seen as a non-invasive biometric that can be collected with or without the speakers' knowledge and transmitted over long distances via telephone lines.

Modern day ASR systems are divided into two classes depending on their desired function: Automatic Speaker Identification (ASI) and Automatic Speaker Verification (ASV). ASI systems attempt to answer the question "who are you?", while ASV systems ask the question "are you who you claim to be?" [2]. An ASV system decides on the identity claim made by the speaker and the output of the system is in the form of a binary result, accept or deny. An ASI system returns the identity of the most likely speaker, from those enrolled in the system.

Replicating published results can be very difficult if authors have not explicitly defined both their system and the conditions under which it was tested. In this paper we aim to clearly define how our ASR system was put together and what conditions our testing was conducted. The remainder of this paper is set out as follows: Section 2 breifly explains the function of an ASR system. Section 3 explicitly explains the databases we used during our simulations

Section 4 explains how our experiments were performed. Finally, section 5 illustrates the result achieved by our GMM based ASR system using various well-known databases.

II. Speaker Recognition

The modern day ASV/ASI system consists of six key components: filtering and A/D, silence removal, front-end processing, pattern matching, decision logic, and enrollment (see Fig. I). The filtering and A/D section is responsible for capturing speech from the real world. The silence is then removed from the speech and converted into a series of highly representative short-time spectral features that highlight the speaker specific properties present in the speech. Using these features the pattern matching section relates them to stored models and calculates a distortion/probability for each model. Using the result of the pattern-matching section the system makes a decision on the validity of the speaker's claim, or the identity of the speaker. However, the system must first be trained to identify speakers, a process commonly referred to as enrollment.

III. Databases

The problem of adequately acquiring speech to train and test an ASR system is overcome through the use of a prerecorded speech database. By using a popular or readily available database results can be directly compared with those previously published by others. For the application of speaker recognition there exists many readily available databases such as YOHO, TIMIT, and ANDOSL.

A. TIMIT / NTIMIT

TIMIT (Texas Instruments Massachusetts Institute of Technology) database allows identification to be done under almost ideal conditions. Therefore, any recognition errors that occur should only be caused by overlapping speaker distributions [6]. The TIMIT database consists of 630 speakers, 70 % male and 30 % female from 10 different dialect regions in America. Each speaker has approximately 30 seconds of speech spread over ten utterances. The speech was recorded using a high quality microphone in a sound proof booth at a sampling frequency of 16 kHz, with no session interval between recordings.

The speech is designed to have a rich phonetic content, which consists of 2 dialect sentences (SA), 450 phonetically compact sentences (SX) and 1890 phonetically diverse sentences (SI). The dialect sentences developed by SRI are spoken by all speakers and were designed to show the variability introduced by the different dialects of the speakers. The phonetically compact sentences were designed by MIT and their purpose was to provide a good coverage of phoneme pairs. Each speaker reads five of these sentences and each sentence is read by seven speakers. The speakers spoke three phonetically diverse sentences that were directly acquired from an existing text sources - Brown Corpus and the Playwrights dialog.

NTIMIT consists of exactly the same speech as TIMIT that has been passed through a local or long distance telephone loop. Through the use of an "*artificial mouth*", each sentence was directly coupled to a carbon button telephone. The speech was then relayed to a local or long distance central office where it was looped back and recorded. The NTIMIT database can be considered to be TIMIT speech suffering from a degradation due to carbon button transducers and actual telephone line conditions.

B. ANDOSL

The ANDOSL (Australian National Database of Spoken Language) is a speech database jointly developed by the Australian National University, the University of Sydney, Macquarie University and the National Acoustic Laboratories, consisting of a few significant diverse phonological groups within Australia. The goal of ANDOSL was to represent as many significant speaker groups within the Australian population as possible [1]. ANDOSL consists of 129 speakers 67 female and 62 male, from the three varieties of Australian English; Broad, General and Cultivated.

Each speaker performed 4 speaking tasks consisting of : 200 phonetically rich SCRIBE sentences, a set of spoken digits, a set

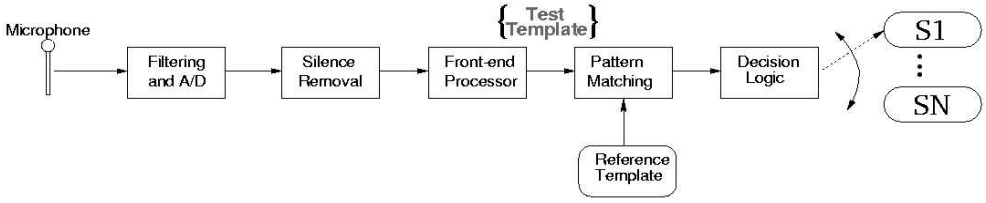

Fig .I. The Basic Components of an ASI System.

of /hVd/ frame words, 2 spontaneous speech tasks. The phonetically rich sentences are contained within the wav subdirectory of each speakers directory denoted by a three digit number preceded by an 's'. ANDOSL has been recored under extremely clean condition at a rate of 20kHz with a resolution of 16 bits.

C. YOHO

The YOHO Database consists of 138 speakers, 108 of them male and 30 female. The data was collected over a three month period, with approximately 3 day verification intervals. The speech data consists of a series of combination-lock phrases, for example 24-52-78. There are 4 enrollment sessions per speaker, each containing 24 phrases and also contains 10 verification sessions per speakers, each containing 4 phrases. The data was recorded at 8kHz with a 3.8kHz bandwidth at 16 bits per sample.

D. Digit-SPL

Digit-SPL is a multi-session database consisting of both male and female speakers. The database was developed at Griffith University in the early part of 2001 and consists of relatively clean speech (an average SNR of 41.6dB) from 68 males and 19 females, spoken on three separate sessions. The sessions are separated by approximately 4-8 weeks. All three session contain ten utterances of two continuously spoken random sequences of five digit numbers, where each digit appeared only once per utterance. The first session contains an additional five repetitions of the isolated word set:"zero", "one", "two", "three", "four", "five", "six", "seven", "eight", "nine".

IV. Speaker Modeling and Classification

A Gaussian Mixture Model (GMM) based text-independent speaker verification and identification system was used to test the discriminate capabilities of the short-time spectral features. This system was similar to the one proposed by Reynolds [4], [5], [6]. Given a feature vector (x_t), the mixture density for speaker s is defined by

$$p(x_t|\lambda_s) = \sum_{i=1}^{M} p_i^s b_i^s(x_t), \tag{1}$$

and can be thought of as the weighted linear combination of M Gaussian densities $b_i^s(x_t)$. Each trained speaker is represented by a model, $\lambda_s = \{\mu_i,\ \Sigma_i,\ \rho_i\}$ where $i = 1, \ldots, M$, μ_i, Σ_i, and ρ_i represent the mean, variance and weighting of the i^{th} mixture respectively. Since there are generally 40 significant acoustic classes in speech, a model order of $M = 32$ was chosen. The models are trained using 15 iterations of the Expectation Maximization (EM) algorithm, with an initial model trained using the k-means algorithm in which cluster centroids are perturbed by 25% of the original cluster variance.

Given the short-time feature representation of the utterance $X = \{x_1, \ldots, x_T\}$, the log likelihood of the utterance belonging to the trained speaker s ($P(X|\lambda_s)$) is found by:

$$P(X|\lambda_s) = \sum_{t=1}^{T} \log p(x_t|\lambda_s) \tag{2}$$

Through the use of a background speaker, the speaker likelihoods are normalized for each spoken utterance. The normalized log-likelihood score is calculated for a given test utterances X by,

$$log\bar{P}(X|\lambda_0) = logP(X|\lambda_s) - \frac{1}{B}\sum_{b=1}^{B} \max_{1\le i\le N_s, i\neq 0} logP(X|\lambda_i) \tag{3}$$

Normalization of the likelihood scores has no effect on the identification performance of the system and are there only to aid in the verification task. Verification of speakers is achieved by applying an experimental threshold to the log likelihood of the trained speaker. The speaker's claim is therefore accepted only if $P(X|\lambda_s)$ exceeds the threshold, i.e.

$$\begin{aligned} if\ P(X|\lambda_s) \ge T_{experimental} \quad & ACCEPT \\ else \quad & DENY \end{aligned} \tag{4}$$

Identification of speakers is implemented using a maximum likelihood classification rule. The speaker's identity is defined by the model that produced the maximum probability, i.e.

$$i^* = arg \max_{1\le i\le N_s} P(X|\lambda_i), \tag{5}$$

where N_s is the total number of trained speakers. The identification error rate (IER) is a measure of how well an identification system can identify speakers. It is simply defined as

$$IER = \frac{N_{ii}}{N_{ti}}, \tag{6}$$

where N_{ii} is the number of incorrect identifications and N_{ti} is the total number of identifications performed. A system can be assumed to be good at identifying speakers if the IER is relatively low (approaching zero). On the other hand if the IER approached one then the system performs poorly in an identification role. The ability of an identification system relies heavily on the number of speakers in the enrolled population. As the number of enrolled speakers increases, the ability of the system to differentiate speakers decreases. This results in an increase in the IER.

The Equal Error Rate (EER) is used for evaluating the performance of a verification system. A general verification system has two possible errors that can be made by the system; an unauthorized speaker can be accepted by the system (false acceptance), or a true speaker can be rejected by the system (false rejection). The false rejection (P_{fr}) and false acceptance (P_{fa}) rates are defined as,

$$P(fr) = \frac{N_{fr}}{N_C}, \quad P(fa) = \frac{N_{fa}}{N_I}, \tag{7}$$

where N_{fr} is the number of times a claimant is rejected by the system, N_C is the number of true claimant tests, N_{fa} is the number of times an impostor is accepted by the system, and N_I is the number

of impostor tests performed. The Error Error Rate (EER) is defined as the point at which these two errors are equal, i.e.

$$EER = P(fr) = P(fa), \tag{8}$$

There exists many methods for finding the EER, a quick and accurate method being the overlapping pdf method. A probability distribution function (pdf) is generated for both P_{fa} and P_{fr} with respect to the experimental threshold value. The region of overlap between these two pdfs is exhaustively searched for the minimum point of equality. If a point does not exist the EER will be assigned the average of both P_{fr} and P_{fa} at their closest value.

V. Experiments

Great care has been taken to explicitly define the way in which each experiment was conducted. This section outlines in great detail how the MFCC features were generated and the speaker models were created using these features. Also the way in which the experiments were conducted is also outlined.

A. Feature Extraction

The feature extraction component of the ASR system was performed using HCopy from the Hidden Markov Toolkit version 3.2 (HTK). The MFCC feature was generated using 24 filters applied to the magnitude spectrum of a 20ms speech frame updated every 10ms. The speech frame was preemphasised using a preemphasis coefficient of 0.95 prior to being windowed by a Hamming window. The resulting MFCC feature was then liftered using sinusoidal liftering with a liftering coefficient of 22. The energy coefficient was in fact the RMS energy of the frame and not in fact c_0. The Delta and Acceleration coefficients were calculated using an orthoganal polynomial fit [7], and the boundary taken care of using simple difference. In the experiments involving Cepstral Mean Subtraction, Cepstrals Mean Subtraction was applied across the complete 39 dimension feature vector.

B. Model Generation

Given the training vectors associated with a particular speaker, the speakers' model was generated as follows. Given the training vectors, a primary GMM model was generated using the k-means algorithm. The entire collection of vectors were segmented into 32 clusters in a linear fashion, i.e. 1 to 2, 2 to 3, etc. The cluster with the greatest variance was chosen as the next candidate for splitting, and the centroids were perturbed by 25% of the clusters variance along each dimension. After a cluster was split in two, all vectors across all clusters were redistributed and all of the centroids updated, a process that was repeated until the total cumulative error across all the vectors no longer changed. When 32 clusters were created the GMM model parameters were extracted. The initial GMM model was then further refined by ten iterations of the EM algorithm.

C. TIMIT Experiments

The speaker models were created using all the sx and si wavefiles in each speakers directory, concatenated into a single 24 second utterance. The two remaining sa wavefiles were used as two independent test segments. The entire 630 speakers included in both the test and train directories were used during the *TIMIT(630)* trials and only the testing directory of the relevant TIMIT database was used during the *TIMIT(168)* tests. The TIMIT8 results refer to an 8kHz down-sampled version of the corresponding 16kHz original database.

D. ANDOSL Experiments

The method of training and testing the ASR system using the ANDOSL database was based on the experiments outlined in [3]. For each speaker a speaker model was created using the first ten phonetically rich sentences contained in the wav subdirectory (001 - 010) joined to make a single utterance. The remaining 190 phonetically rich sentences (011 - 200) were used as 190 independent test cases.

E. YOHO Experiments

There are two published methods of using YOHO for verification, the first requires building each speaker model using only sessions 1 - 3 and using the fourth to build cohort models [2]. The second is to use all of the four enrollment session (sessions 1 - 4) to generate speaker models. Both of these method were used in our experiment, they are labeled YOHO-4 and YOHO-3 in which four and three enrollment sessions were used to train the relevant speaker model. All of the speakers verification sessions were used to test the speaker models in both cases.

F. DIGIT-SPL Experiments

The isolated word set from session 1 was used to generate speaker models for all 87 members of the database. The speaker models were tested using the 30 continuous utterances contained in sessions 1, 2, and 3, labeled in our results as DIGIT-SPL1, DIGIT-SPL2, and DIGIT-SPL3 respectively.

VI. Results

The results of the experiments conducted are shown in Tables I and II. The identification error rate and average ranking of incorrectly identified speakers are included, not to mention the equal error rate when the system is used in a verification role.

From Table I it can be seen that including the delta and acceleration coefficients has a negative affect on ASR performance when a database consisting of clean broadband speech is used. The use of deltas and acceleration coefficients has a positive effect when used on telephone speech as can be seen in the NTIMIT and YOHO results. In the presence of inter-sessional variability such as in the DIGIT-SPL tests, the use of delta or acceleration coefficients also had a negative effect.

VII. Conclusion

Through the experiments performed on the included readily available databases the following points can be made. Due to the diverse range of speakers contained in the ANDOSL database and the vast amount of data used to train each speaker. The ANDOSL database poses no great challenge to an ASR system, with our ASR system achieving an almost perfect performance in both verification and identification roles. The TIMIT database has shown its usefulness for testing the ability of an ASR system to separate speakers under ideal clean conditions. It has also been shown that using the MFCC feature with delta, acceleration, and energy coefficients has a negative effect on clean speech a trend worth looking into regarding LPCC coefficients.

References

[1] Millnar J. B., VonWiller J. P., Harrington J. M., Dermody P. J., "The Austrakian national database of spoken language," in *Proc. Inter. Conf. on Acoustics, Speech & Signal Processing* (ICASSP'94), Vol. 1, pp. 97 - 101, 1994.

[2] Campbell J. P., "Speaker recognition: a tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437-1462, 1997.

Database Used	MFCC			$(E + MFCC) + \Delta + \Delta^2$			$(MFCC) + \Delta + \Delta^2$		
	IER	EER	Avg. Rank	IER	EER	Avg. Rank	IER	EER	Avg. Rank
TIMIT(168)	0.0000	0.0036	NA	0.0033	0.0021	2.00	0.0033	0.0021	2.00
TIMIT(630)	0.0097	0.0073	3.75	0.0211	0.0040	2.92	0.0236	0.0049	3.07
NTIMIT(168)	0.6478	0.1335	15.35	0.5863	0.1215	14.7	0.4723	0.1063	14.17
NTIMIT(630)	0.5556	0.0875	24.15	0.7181	0.1254	43.49	0.7002	0.1105	39.15
TIMIT8(168)	0.0345	0.0124	2.63	0.0684	0.0092	2.62	0.06514	0.0100	2.45
TIMIT8(630)	0.1056	0.0194	4.98	0.1462	0.0171	5.24	0.1218	0.0187	5.3
NTIMIT8(168)	0.5656	0.1236	16.16	0.4951	0.1003	12.76	0.5733	0.1135	13.38
NTIMIT8(630)	0.6320	0.1060	38.72	0.6223	0.0926	31.1	0.6011	0.0865	30.77
ANDOSL-20K	0.0252	0.0060	2.89	0.0017	0.0007	2.76	0.00044	0.00039	3.00
ANDOSL-8K	0.00098	0.0024	4.75	0.0037	0.0021	2.91	0.00112	0.00141	2.56
YOHO-4	0.1283	0.0397	6.22	0.0517	0.0143	7.11	0.0355	0.0127	6.37
YOHO-3	0.1607	0.0445	6.88	0.0706	0.0177	7.08	0.0541	0.0151	6.15
DIGIT-SPL-1	0.0151	0.0735	4.15	0.04497	0.0761	8.68	0.0318	0.0683	4.52
DIGIT-SPL-2	0.2058	0.0957	8.07	0.2975	0.1152	8.92	0.2131	0.0998	7.07
DIGIT-SPL-3	0.2604	0.1104	9.68	0.3152	0.1335	12.25	0.2752	0.1149	9.37

TABLE I

RESULTS ACQUIRED USING THE GMM BASED ASR SYSTEM.

Database Used	$(E + MFCC) + \Delta + \Delta^2 + CMS$		
	IER	EER	Avg. Rank
TIMIT(168)	0.02606	0.00652	3.75
TIMIT(630)	0.10154	0.01379	5.11
NTIMIT(168)	0.62541	0.12412	15.47
NTIMIT(630)	0.71812	0.10327	33.62
TIMIT8(168)	0.18387	0.03306	5.84
TIMIT8(630)	0.337937	0.03822	12.50
NTIMIT8(168)	0.53074	0.09378	14.54
NTIMIT8(630)	0.74411	0.12563	43.93
YOHO-4	0.04461	0.01981	8.57
YOHO-3	0.05766	0.02036	7.70
DIGIT-SPL-1	0.06551	0.02188	10.49
DIGIT-SPL-2	0.34368	0.08840	9.08
DIGIT-SPL-3	0.38851	0.11423	11.75

TABLE II

RESULTS ACQUIRED USING THE GMM BASED ASR SYSTEM WITH CMS.

[3] Tran D. and Wagner M., "A proposed likelihood transformation for speaker verification," in *Proc. Inter. Conf. on Acoustics, Speech & Signal Processing* (ICASSP'00), pp. 1069 - 1072, 2000.

[4] Reynolds D. A. and Rose R. C., "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, vol. SAP-3, no. 1, pp. 72-83, January 1995.

[5] Reynolds, D. A., "A Gaussian mixture modeling approach to text-independent speaker identification," Ph.D. Thesis, Georgia Institute of Technology, 1992.

[6] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1, pp. 91-108, 1995.

[7] F. K. Soong and A. E. Rosenberg, "Use of instantaneous and transitional spectral information in speaker recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-36, no. 6, pp. 871-879, 1988.