Journal of VLSI Signal Processing 36, 91–104, 2004

Acoustic Feature Analysis and Discriminative Modeling of Filled Pauses for Spontaneous Speech Recognition

CHUNG-HSIEN WU AND GWO-LANG YAN
Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, Republic of China

Received October 30, 2001; Revised June 5, 2002; Accepted September 3, 2002

Abstract. Most automatic speech recognizers (ASRs) concentrate on read speech, which is different from spontaneous speech with disfluencies. ASRs cannot deal with speech with a high rate of disfluencies such as filled pauses, repetitions, lengthening, repairs, false starts and silence pauses. In this paper, we focus on the feature analysis and modeling of the filled pauses "ah," "ung," "um," "em," and "hem" in spontaneous speech. Karhunen-Loéve transform (KLT) and linear discriminant analysis (LDA) were adopted to select discriminant features for filled pause detection. In order to suitably determine the number of discriminant features, Bartlett hypothesis testing was adopted. Twenty-six features were selected using Bartlett hypothesis testing. Gaussian mixture models (GMMs), trained with a gradient decent algorithm, were used to improve the filled pause detection performance. The experimental results show that the filled pause detection rates using KLT and LDA were 84.4% and 86.8%, respectively. A significant improvement was obtained in the filled pause detection rate using the discriminative GMM with KLT and LDA. In addition, the LDA features outperformed the KLT features in the detection of filled pauses.

Keywords: filled pause, disfluency, Guassian mixture model, speech recognition, Karhunen-Loéve transform, linear discriminant analysis

1. Introduction

The recent growing demand for automatic speech recognizers (ASRs) has been manifested in applications such as dialog systems, call managers and weather forecasting systems. The most noticeable problem for these systems is the poor recognition rate for filled pauses because spontaneous speech is punctuated with and interrupted by a wide variety of seemingly meaningless words. This disfluent speech contains filled pauses, repetitions, lengthening, repairs, false starts, and silence pauses. All kinds of disfluencies generally destroy the smooth speaking style of speech and therefore degrade the performance of ASR. Past work on the detection of disfluent speech was based on acoustic, language, and prosodic information models. In the acoustic model approach, most of the research treated the disfluency as a general resembling recognition model [1, 2]. However, these speech recognizers are typical HMM based and accept only fluently read or planned speech without disfluencies. These works suffer difficulties in dealing with filled pauses and word lengthening because the duration of a phone tends to lengthen differently compared to general speech. For the language model-based approach, some previous research [3–5] tried to correct the recognition errors caused by disfluency using language models. These works either took the difluency into account or skipped the disfluency in the language model. This is not effective for dealing with filled pauses because they can be inserted at arbitrary positions. The corpus-based language models cannot completely model all of the various kinds of disfluent conditions. The skip strategy cannot preserve the spontaneous speech structure from destruction because it ignores the important roles of disfluent words [4, 6, 7]. Other research [8, 9]

analyzed the prosody of disfluent speech. Prosodic cues such as pitch and duration were exploited to derive rules and features to detect the disfluent positions in speech. These rules and features using only pitch and duration are still not enough to model all of the types of filled pauses.

In this paper, the filled pauses "ah," "ung," "um," "em," and "hem" are investigated. The approach used for detecting these pauses is based upon two principles. The first principle was to prevent ASR performance degradation due to the presence of filled pauses. The second important principle was that filled pauses play valuable roles, such as markers of discourse structure [7], thinking and helping a speaker maintain his turn in the conversation and meaningful oral communication. Filled pauses can be characterized by two acoustic properties according to our investigation. These properties are the nasal effect and lengthening. Many salient features for describing these phenomena have been analyzed and proposed in [10–14]. For example, nasal sounds are characterized by the first two formant frequencies (at about 300 and 1000 Hz). The mean magnitude difference between the amplitudes of formant 1 and antiformant 1 is greater for the oral vowel than the nasal vowel [10, 11]. The distinctive spectral traits of nasal consonants are low first formant with higher intensity than the upper formants and a set of weak formants at 300–4000 Hz [12, 13]. The magnitude ratio of formant 2 to formant 1 is smaller in the nasal consonant [14]. Lengthening is characterized by a steady spectrogram. The spectral/cepstral coefficients used to model the vocal tract are chosen as the features for detection. In our approach, 48 features that convey the properties of the filled pauses described above are extracted first. The Karhunen-Loéve transform (KLT) and linear discriminant analysis (LDA) are then performed on these features. The Bartlett hypothesis testing [15] was adopted to determine the number of discriminant features. Twenty-six features were selected using the Bartlett hypothesis testing under a confidence level 97.5%. Using these selected features, the filled pause detection rate was about 76.8% for the KLT features and 78.1% for the LDA features. To increase the discriminability, these features were modeled further using Gaussian mixture models (GMMs) [16], whose weights are estimated using a gradient decent training algorithm that recursively minimizes the detection error rate. An optimal threshold is set to achieve the best detection rate.

The proposed architecture for filled pause detection is shown in Fig. 1. In the training process, the speech features are extracted and then analyzed using the Karhunen-Loéve transform and linear discriminant analysis. The feature number was determined using the Bartlett hypothesis testing. These features were then modeled using the filled pause GMMs to form a model database. Similarly, the same features for the normal or fluent speech were modeled using a fluency GMM. To increase the discriminability, the GMMs were trained using a discriminative algorithm to optimize the detection rate. In the filled pause detection process, a threshold is then defined to detect the filled pauses by the verification scores calculated from the GMMs.

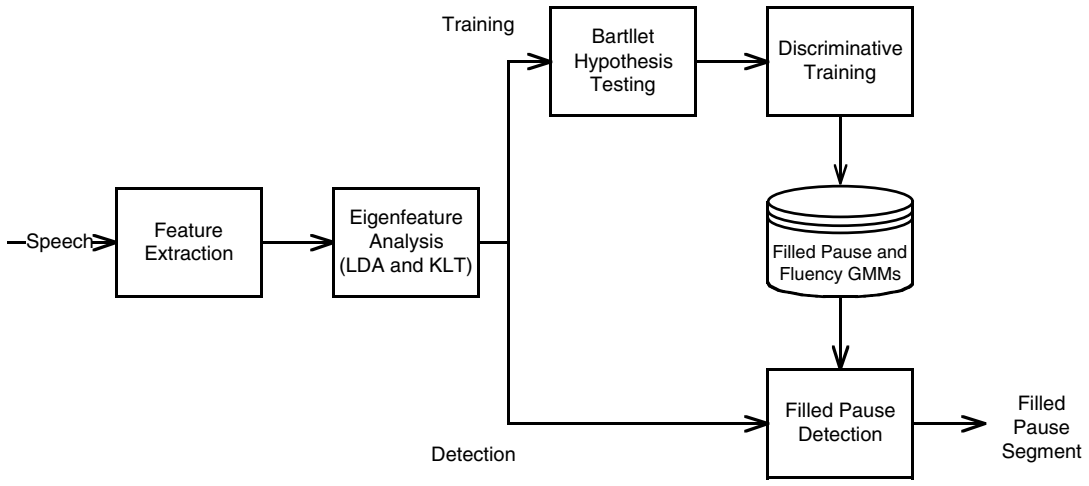


Figure 1. System architecture for the detection of filled pauses.

This paper is organized as follows. In Section 2, we describe the features chosen for characterizing the filled pauses. The Karhunen-Loéve transform and linear discriminant analysis methods are also described. In Section 3, the architecture and the GMM discriminative training are depicted. The experimental results are shown in Section 4. A brief conclusion is presented in Section 5.

2. Feature Analysis of Filled Pauses

Because filled pauses may appear anywhere in an utterance when people are talking with each other, a database of spontaneous conversations was collected from several spontaneous conversations. The spontaneous speech database contains over 8 hours of recorded speech, spoken by over 40 speakers of both sexes. These spontaneous speech utterances were transcribed, segmented and tagged into 2,160 sentences. According to our preliminary observation from the database, the filled pauses can be summarized as having two properties: the nasal effect and lengthening properties. The feature analysis and selections for these two properties are described in the following sections.

2.1. *Lengthening Property*

All filled pauses in spontaneous speech have a common property: lengthening. For voice lengthening, the vocal cord vibrates periodically and the vocal tract is maintained in a relatively stable configuration throughout the utterance. In other words, the produced voice changes smoothly. Figures 2(a) and (b) show the waveform and spectrogram of the speech utterance "嗯…你好" (um . . . how are you) in Mandarin. The lengthening voice "um" occurs at the beginning of the utterance. The spectrogram is nearly steady compared to the voice in the middle of the utterance. Two kinds of spectral/cepstral coefficients are employed to characterize this property. The first kind of coefficient, which is famous in modeling the vocal tract, is the linear predictive coding (LPC) coefficient. The second feature, which considers the human hearing perception and has been proven useful for speech recognition, is the mel-frequency cesptrum coefficient (MFCC). We choose 12 MFCCs, 12 delta MFCCs, and 12 LPCs as the features used to detect and analyze the lengthening property. Figure 2(c) shows an example of these features, in which 12 MFCCs and 12 delta MFCCs are stable in the lengthening part.

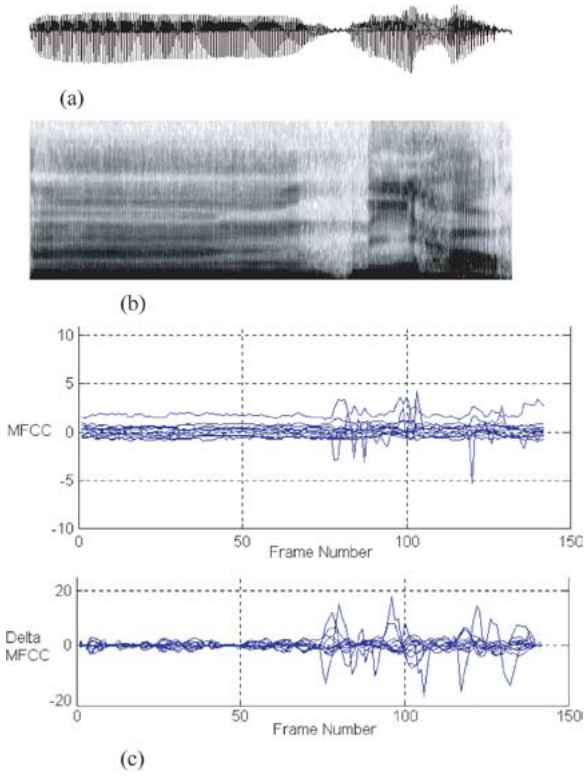


Figure 2. (a) Waveform, (b) spectrogram and (c) 12 MFCCs and 12 delta MFCCs of the utterance "嗯…你好" (um . . . How are you).

2.2. *Nasal Effect Property*

The second property in filled pauses is the nasal effect. This property is found in the filled pauses "ah," "ung," "um," "em," and "hem." In the production of nasal sounds, the resonance characteristics are conditioned by the oral cavity characteristics forward and backward from the velum and by the nasal tract characteristics from the velum to the nostrils. A special production procedure causes this particular formant change. Some research [10, 11] has reported on nasalized voices. The noticeable cues are the first two formants (at about 300 and 1000 Hz). The first two formant frequencies for normal sounds are generally at about 250–800 and 700–2500 Hz. Figure 3 shows an example of "a" and its nasalized sound "ah." We can see that the formant frequencies of F1 and F2 for /ah/ are changed to about 250–800 and 700–2500 Hz compared to the oral sound /a/. Figure 4 shows the distributions of F1 and F2 for the vowels "a," "i," "u," "e," and "o," and the filled pauses "ah," "ung," "um," "em," and "hem." The "x" marks, representing the filled pauses "ah," "ung," "um," "em,"

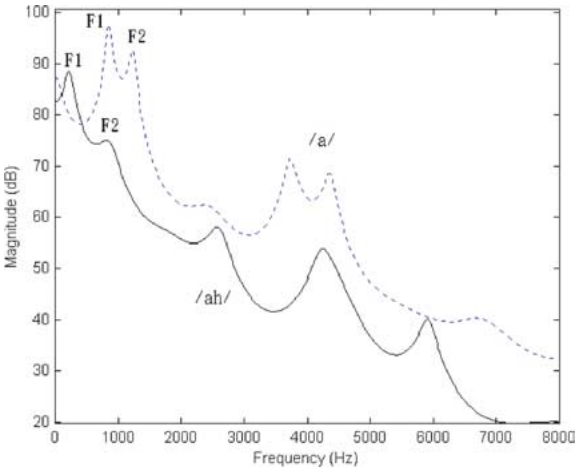


Figure 3. Spectral envelopes of /a/ (dash) and /ah/ (solid).

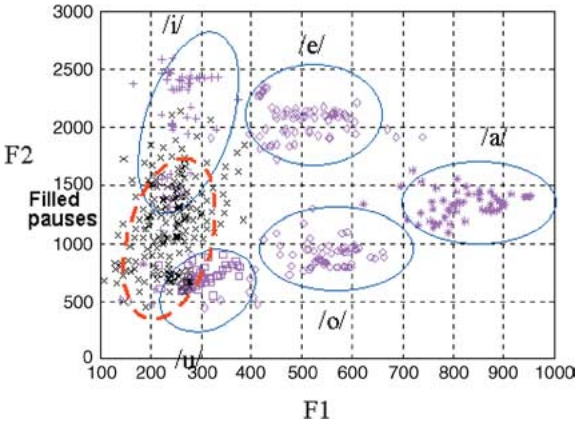


Figure 4. Plot of F2 versus F1 for vowels "a," "i," "u," "e," and "o," and filled pauses "ah," "ung," "um," "em," and "hem."

and "hem" in this figure, can be distinguished from the vowels in the vowel triangle [10]. The formant 1 (F1) and formant 2 (F2) frequencies for "ah," "ung," "um," "em," and "hem" are useful for characterizing the nasal sounds. They were also chosen as the features for filled pause detection. We also include formant 3 (F3) for analysis. In this paper, ARMA framework with modified recursive least square algorithm [17] is applied to explore the spectral information of a time varying signal and track adaptively and directly the polynomial zeros and poles of the transfer function. The average frequencies of F1, F2, and F3 for "ah," "ung," "um," "em," and "hem" are listed in Table 1 calculated from our corpus described in Section 5.1.

The second cue for nasal sounds is the mean magnitude difference between the amplitudes of F1 and

Table 1. The formant frequencies F1 and F2 of "ah," "ung," "um," "em" and "hem".

	ah	ung	um	em	hem
Formant 1	233 Hz	239 Hz	245 Hz	191 Hz	203 Hz
Formant 2	1268 Hz	1021 Hz	1253 Hz	720 Hz	778 Hz

Table 2. Average mean magnitude differences between F1 and Z1 for (a) /a/ compared to their nasalized counterparts /ah/ and (b) /e/ compared to its nasalized counterpart /em/.

	(a)		(b)	
	/a/	/ah/	/e/	/em/
Mean magnitude difference (db)	17.11	8.87	14.74	9.56

antiformant 1 (Z1) frequencies. If we denote the spectrum envelope function as $SE(f)$ where f is the input frequency, the mean magnitude difference is defined as

$$MD(F1, Z1) = SE(F1) - SE(Z1) \tag{1}$$

In Table 2, the average mean magnitude differences for /a/ compared to its nasalized counterpart /ah/ and /e/ compared to its nasalized counterpart /em/ are estimated and compared. It is clear that the mean magnitude difference for oral vowel is greater than that for nasal vowel [10, 11]. To include more features for further analysis, $MD(F1, Z1)$, $MD(F2, Z2)$, and $MD(F3, Z3)$ are also taken into account.

The third cue for the nasal effect in filled pauses generally occurs in nasal consonants such as "ung" and "um" [11–13]. The characteristic of the nasal consonant is its distinctive spectral traits. According to the reports, the distinctive spectral traits of nasal consonants are (1) low first formant with higher intensity than the upper formants and (2) low amplitudes for the upper formants. The properties described above can be observed in two parts. The first is the formant magnitude ratio, which is defined in the following.

$$FMR(F2, F1) = \frac{SE(F2)}{SE(F1)}, \tag{2}$$

where $SE(F1)$ represents the magnitude of F1. The equation formulates the degree of the decrease in magnitude when the voice is produced through the nostrils. Figure 5 shows a histogram of the formant magnitude ratio for "ung" and "um." The mean of this distribution is about 0.08. Figure 6 shows a histogram of the

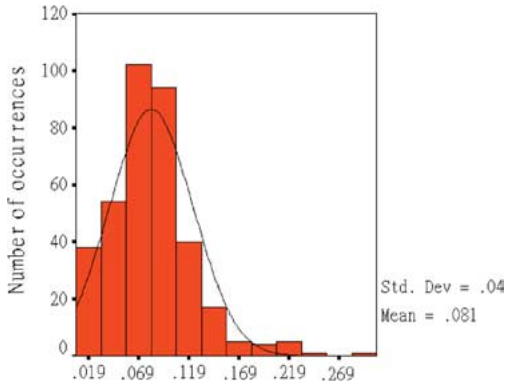


Figure 5. Histogram of formant magnitude ratio for "ung" and "um".

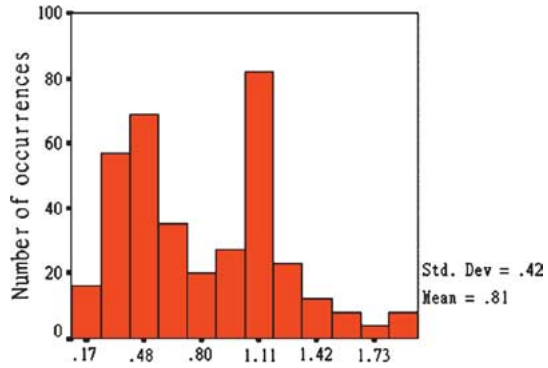


Figure 6. Histogram of formant magnitude ratio for normal voices.

formant magnitude ratio for oral voices. The distribution mean is about 0.8. The formant magnitude ratio is very special for the nasal characteristics of "ung" and "um" and can be used to characterize the property "low first formant with higher intensity than the upper formants." Similarly, the features $FMR(F2, F1)$, $FMR(F3, F1)$ and $FMR(F2, F3)$ were also included for analysis.

The fourth cue for the nasal effect is the low amplitudes for the upper formants. This also describes the dramatic decrease in magnitude when the voice is produced through the nostrils. Therefore, $SE(F1)$, $SE(F2)$ and $SE(F3)$ were also chosen as features for analysis.

3. Discriminant Feature Analysis and Selection

According to our previous analysis, there are 48 features that are possibly useful for filled pause detection. They are 12 MFCCs, 12 delta MFCCs, 12 LPCs, F1, F2, F3, $MD(F1, Z1)$, $MD(F2, Z2)$, $MD(F3, Z3)$, FMR (F2, F1), FMR (F3, F1), FMR (F3, F2), SE(F1), SE(F2), and SE(F3). In practice, however, this number of features is too large to allow robust and fast detection. It is desirable to obtain a suitable set of features to achieve an acceptable performance. A common way to resolve this problem is to use dimensionality reduction techniques and discriminant analysis. Two of the most popular techniques for this purpose are: Karhunen-Loéve transform and linear discriminant analysis. In the following sections, we will describe how these two techniques were applied to the filled pause detection problem and how to determine a suitable feature number m to achieve an acceptable result.

3.1. Karhunen-Loéve Transform

Efficient salient feature selection is an important issue in classification. The Karhunen-Loéve transform (KLT), also known as the principal component analysis or "eigenfeatures," was adopted to choose the most efficient features for face recognition and image retrieval [18, 19]. In this approach, KLT is employed to remove information redundancies and generate orthogonal features so that the selected features are less confusing and more discriminant. Given an n-dimensional vector X, it can be expanded and approximated by

$$X = VY \tag{3}$$

where the columns of the $n \times n$ square matrix V are an orthogonal basis vector with $V^T V = I$; Y is a transformed feature vector of the random vector X. In order to derive the efficient features for discrimination, the approximation of X using $m < n$ columns of V gives

$$\hat{X}(m) = \sum_{i=1}^{m} y_i v_i \tag{4}$$

where v_i $(i = 1..m)$ are the column vectors of V. In other words, this is the projection of X onto the subspace spanned by the m orthonormal eigenvectors. The approximation estimation can be defined according to the mean-square error as

$$Q(X, \hat{X}) = E[\|X - \hat{X}(m)\|^2]. \tag{5}$$

Our goal now is to choose eigenvectors that minimize the mean-square error Q. From Eq. (5) and taking into

account the orthonormality property of the eigenvectors, we have

$$E[\|X - \hat{X}(m)\|^2] = \left[\left\|\sum_{i=m+1}^{n} y_i a_i\right\|^2\right] = E\left[\sum_i \sum_j \left(y_i v_i^T\right)(y_j v_j)\right] = \sum_{i=m+1}^{n} v_i^T E[XX^T] v_i \quad (6)$$

Combining Eq. (6) and the eigenvector definition, we finally get

$$E[\|X - \hat{X}(m)\|^2] = \sum_{i=m+1}^{n} v_i^T \lambda_i v_i = \sum_{i=m+1}^{n} \lambda_i \quad (7)$$

According to the result from Eq. (7), the mean-square error is proportional to the summation of the smallest eigenvalues from $m+1$ to n. Consequently, the chosen $v_1, v_2, .., v_m$ eigenvectors are associated with the m largest eigenvalues in the covariance matrix of X, defined as

$$\sum_X = E[(X - \mu_X)(X - \mu_X)^T], \quad (8)$$

where μ_X is the mean vector of X. The transformed features $y_1, y_2, .., y_m$ are then computed as follows.

$$y_i = v_i^T (X - \mu_X), \quad i = 1, 2, .., m. \quad (9)$$

These features are the optimal features for producing the minimum mean-square error. We call these features the KLT features, employed for the acoustic discriminative modeling of filled pauses in the following experiments. In this approach, several sets of discriminant features such as MFCCs, LPCs, formant frequencies are not in a same magnitude, so we adopt correlation matrix instead of covariance matrix defined in Eq. (8) before applying the KLT or LDA. The correlation matrix is defined as:

$$R_X = E\left[\left(\frac{X - \mu_X}{\sigma_X}\right)\left(\frac{X - \mu_X}{\sigma_X}\right)^T\right] \quad (10)$$

where σ_X is the standard deviation vector of X.

3.2. Linear Discriminative Analysis

Linear discriminant analysis (LDA) is another famous feature selection method [19]. It has been widely used to search for those vectors in the underlying space that best discriminate among classes. In other words, LDA creates a linear transform of these features to maximize the largest mean differences between the desired classes. The following matrices for the random vector X are defined as

1. Within-class scatter matrix:

$$S_w = \sum_{i=1}^{C} P_i E\left[\left(X_i - \mu_{X_i}\right)\left(X_i - \mu_{X_i}\right)^T\right] \quad (11)$$

   where X_i is the i-th class of $X(i = 1..C)$, P_i is the a priori probability of class i and $P_i \cong n_i/N$ is computed by the number n_i of samples belonging to class i divided by the total number N.
2. Between-class scatter matrix:

$$S_b = \sum_{i=1}^{C} P_i \left(\mu_{X_i} - \mu_X\right)\left(\mu_{X_i} - \mu_X\right)^T \quad (12)$$

   where μ_{X_i} is the sample mean of class i and μ_X is the global mean vector and accumulated using

$$\mu_X = \sum_{i=1}^{N} P_i \mu_{X_i} \quad (13)$$

The goal is to maximize the between-class measure while minimizing the within-class measure. A number of different criteria were defined using various combinations of these scatter matrices in a "trace" or "determinant" formulation. One way to do this is to maximize the ratio $\frac{\det(S_b)}{\det(S_w)}$ [19]. This ratio has proven that if S_w is a nonsingular matrix, then this ratio is maximized when the column vectors of the projection matrix are the eigenvectors of $S_w^{-1} S_b$ associated with the largest eigenvalues. We call these features the LDA features. Take the filled pause "ung" as an example. In Fig. 7, the projection similarities are introduced to explain the discriminant property between "ung" and not-"ung." The horizontal axis is the projection similarity to the sound "ung" and the vertical axis is the projection similarity to the sound not-"ung." Figure 7(a) shows a distribution of the original features and Figs. 7(b) and (c) show the distributions of the features using KLT and LDA ($m = 26$), respectively. From this figure, the KLT and LDA features give better separability than the original

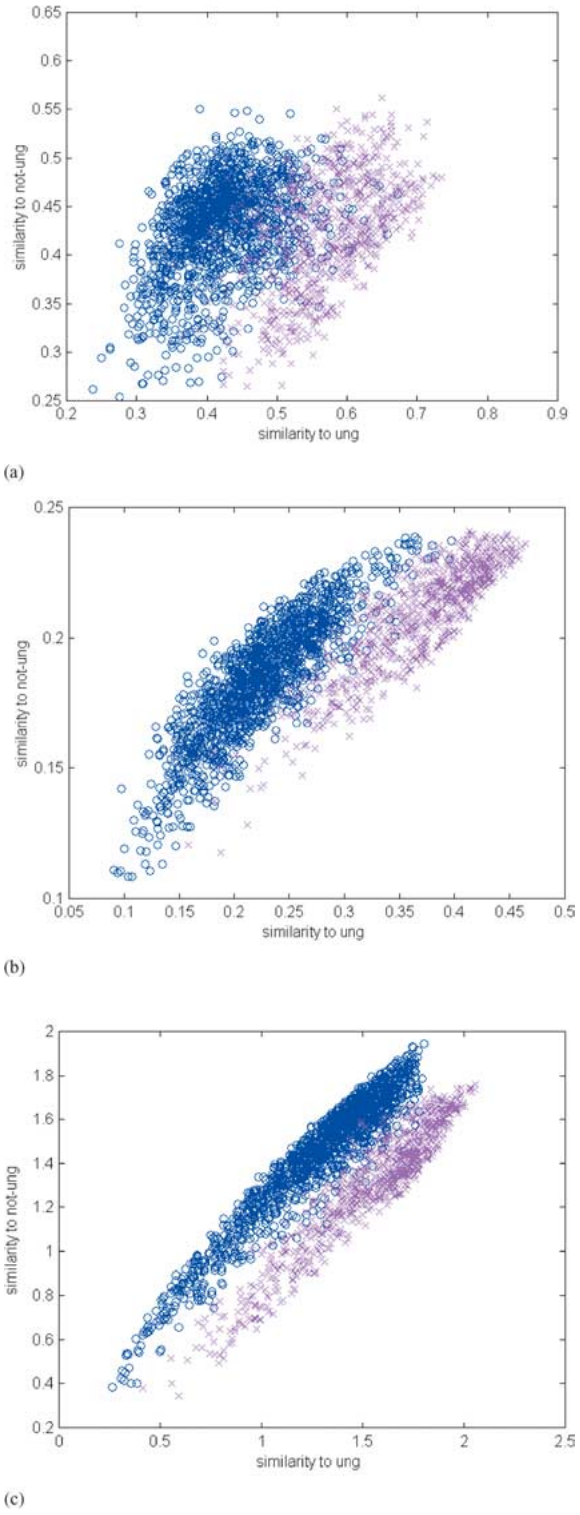


Figure 7. Comparison of the discriminability for "ung" and not-"ung" using (a) original features, (b) KLT features, and (c) LDA features. Circle: Filled pause. Cross: not-filled pause.

features when a suitable feature number is selected. It is because the KLT and LDA projections are likely to achieve better separability for a selected feature number. In our approach, for comparison between the KLT and LDA features, the values of m were chosen as the same.

3.3. *Bartlett Chi-Square Testing*

Determining the feature number m is another important task. In this section, our motivation is based on selecting a suitable m with an acceptable performance without paying too much effort to run all the experiments for m from 1 to 48. The following two methods, conventional and statistical methods, are adopted and compared. The conventional method utilizes the residual error to evaluate the efficiency for different numbers of features. If we rank the eigenvalues of correlation matrix R_X to $\lambda_1, \lambda_2, .., \lambda_n$, according to Eq. (7), the residual error is simply defined according to the sum of the $n - m$ smallest eigenvalues not used, compared to the total sum of eigenvalues. It is written as

$$\text{Residual Error} = \frac{\sum_{i=m+1}^{n} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{14}$$

The residual error is always chosen as 5% thereby the features covering all of the principle component characteristics. This is a tradeoff between the detection rate and computation time. We would like to select a suitable feature number for better discrimination and avoid confusion due to unsuitable features. For formal statistical analysis, Bartlett [15] proposed the chi-square testing method for testing the null hypothesis:

$$H_0 : \lambda_{m+1} = \cdots = \lambda_n \tag{15}$$

The statistical formula is defined as

$$\begin{aligned} \chi^2 &= q\left[-\ln\left|\sum_x\right| + \sum_{j=1}^{k} \ln \lambda_j + m \ln(\ell) \right] \\ &\approx \chi^2_{\frac{1}{2}(n-k-1)(n-k+2)} \end{aligned} \tag{16}$$

where we choose $k = n - m$,

$$q = n - k - \frac{1}{6}\left(2m + 1 + \frac{2}{m}\right), \quad \text{and}$$

$$\ell = \frac{1}{q}\left(\text{tr}\left(\sum_x\right) - \sum_{j=1}^{k} \lambda_j\right).$$

In Bartlett chi-square testing, if the eigenvalues in the higher order are below a threshold and change smoothly (that is, the null hypothesis is accepted), it means that these higher order features are not significant for discrimination and can be ignored. These two methods were adopted in our approach for choosing the most suitable m for performance comparison. The value of "m" was selected as 26 according to Eq. (16) under a confidence level of 97.5%. Although the 26 selected features cover only 91.83% of the important attributes, the best detection rate for filled pause detection was about 76.8% and slightly better than using the features without discriminant analysis. This result was not surprising because using analyzed features is more powerful for precisely expressing and discriminating the difference between filled pauses and fluency. However, the improvement was not significant. In order to describe the specific difference more precisely, a Gaussian mixture model with multi-mixtures, i.e., many sets of analyzed features, was employed for better modeling.

4. Filled Pause Modeling Using Gaussian Mixture Model

4.1. Gaussian Mixture Model

The Gaussian mixture model (GMM) [14, 16] is a commonly used statistical model in speech and speaker recognition. In this model, the covariance matrix is usually assumed to be diagonal in applications. This assumption discards the cross-correlation between parameters and takes advantage of less computation. In speech or speaker recognition systems, the features are modeled as a class whose output probability is represented by a Gaussian mixture density. In the GMM, each Gaussian mixture with its weight of importance is used to calculate the output probability. This is because each Gaussian mixture will have a different contribution to the output probability. The framework is depicted in Fig. 8. This architecture is suitable for modeling the categorical data more precisely because it contains a large number of mixtures with their weights for describing a distribution. In this estimation process, the GMM calculates the probability of a feature $\vec{x}$ using a weighted combination of multi-variate Gaussian densities defined as

$$GMM_\lambda(\vec{x}) = \sum_{i=1}^{D} w_i N_i(\vec{x}) \tag{17}$$

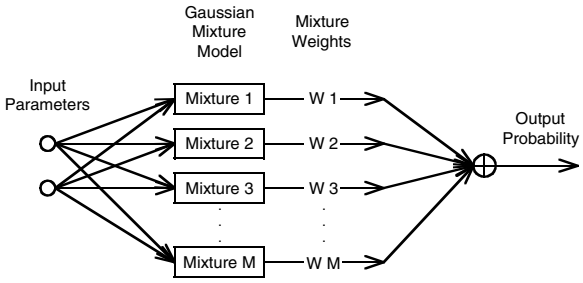


Figure 8. The framework of the GMM.

where $N_i(\vec{x})$ is the multi-variate Gaussian distribution and defined as

$$N_i(\vec{x}) = \frac{1}{\sqrt{2\pi^D |\sum|}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu}_i)^T \sum^{-1} (\vec{x} - \vec{\mu}_i)\right) \tag{18}$$

and w_i is the mixture weight corresponding to the i-th mixture and satisfies

$$\sum_{i=1}^{D} w_i = 1. \tag{19}$$

λ is the model and described by

$$\lambda = \left\{w_i, \vec{\mu}_i, \sum\right\} \tag{20}$$

where $\vec{\mu}_i$ is the mean of the i-th Gaussian mixture and $\sum$ is the diagonal covariance matrix. In our approach, the 26 selected features used to analyze each filled pause in Section 2 were modeled using the GMM with 16 Gaussian mixtures using the modified k-means algorithm [20]. The weights were initially set to the ratio of the number of feature sets belonging to each Gaussian mixture with respect to the total number. They were then updated based on the gradient descent-training algorithm. There were a total of 5 filled pause GMMs for "ah," "ung," "um," "em," and "hem". Similarly, the same features for all fluent speech samples were modeled using a fluency GMM. In this modeling process the input features were fed into all Gaussian mixtures with their weights. The probability with a weighted summation of these probabilities was output using Eq. (17). The output probabilities of filled pause GMMs and fluency GMM are used to discriminate the filled pauses from fluent speech. A threshold T was used to determine the acceptance or rejection of a candidate filled pause.

4.2. Discriminative Training of Mixture Weights

In order to achieve a better discrimination between filled pauses and fluent speech, a verification function was defined to form a linear discriminator whose weights are discriminatively trained. For a given input speech x, the verification function is written as

$$V(x_t; H) = \log\left(\frac{GMM_H(x_t)}{GMM_{\bar{H}}(x_t)}\right) = \log\left(\frac{\sum_m W_{H,m} N_{H,m}(x_t)}{\sum_m W_{\bar{H},m} N_{\bar{H},m}(x_t)}\right) \quad (21)$$

where x_t represents the t-th feature vector of the input speech, and the weight vectors $W_{H,m}$ and $W_{\bar{H},m}$ are the m-th mixture weights for the filled pause and fluency models H and $\bar{H}$, respectively. The terms $W_{H,m}N_{H,m}(x_t)$ and $W_{\bar{H},m}N_{\bar{H},m}(x_t)$ are the output probabilities of the filled pause and fluency models, respectively. A loss function was defined and minimized with respect to the weights. The loss function represents a smooth functional form of the verification score. It takes the form of a sigmoid function, written as

$$R(W_M, x_t) = \frac{1}{1 + \exp[-\eta b V(x_t; H)]} \quad (22)$$

where

$$b = \begin{cases} -1 & \text{if } x_t \in H \\ +1 & \text{if } x_t \in \bar{H} \end{cases} \quad (23)$$

$$W_M = [W_{H,M}, W_{\bar{H},M}] \quad (24)$$

and $\eta(=0.01)$ is a constant that controls the steepness of the sigmoid function.

According to the usual discriminative methodology, an optimization criterion can be defined as the minimization of the loss function and a gradient descent algorithm can interactively update the mixture weights. However, because the probability density function of x_t is not known, a gradient-based iterative procedure is used to minimize R as follows.

$$(W_M)_{n+1} = (W_M)_n - \varepsilon \nabla R((W_M)_n, X) \quad (25)$$

where ε is the updated step size and $\nabla R((W_M)_n, X)$ is the gradient of the loss function with respect to W_M evaluated by the training samples.

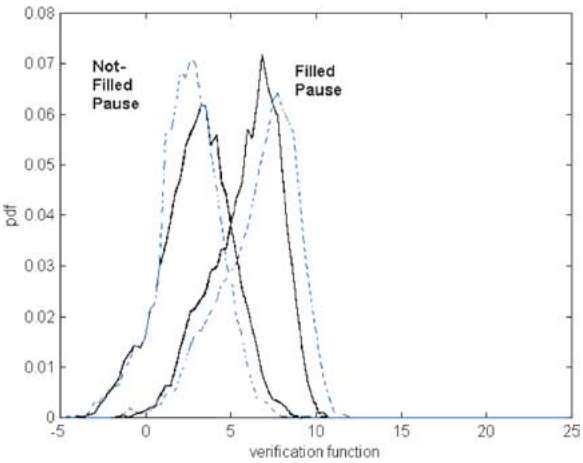


Figure 9. Probability density of the verification scores of LDA for "filled pause" and not-"filled pause." Solid line: before discriminative training. Dotted line: after discriminative training.

In our approach, using Eq. (21), we plotted the verification scores for the "filled pause" models and not-"filled pause" model. The histograms of the verification scores are shown in Fig. 9. The solid histogram on the right represents the distributions of the training samples with filled pauses. Similarly, the solid histogram on the left represents the distributions of the training samples with fluent speech. After discriminative training, the histogram of verification scores is shown in Fig. 9 with dotted line. We can see that the detection error rate for filled pauses is smaller when an optimal decision boundary is chosen after discriminative training.

5. Experiments

5.1. Database Collection

In the following experiments, a database was collected from 40 speakers. They were brought into the lab to answer questions proposed by a questioner. These conversational speech utterances were recorded and tagged into 2,160 sentences. About 12.5% of the database contains filled pauses and their corresponding sentence numbers are listed in Table 3. The most frequently used filled pause is "um" and the least frequently used filled pause is "ung." The position of the filled pause is often at the beginning of a sentence and sometime appears in the middle of a sentence. This database is divided into training and testing databases. The training database

Table 3. The number of sentences for all filled pauses in the database.

	uh	ung	um	em	hem
No. of sentences	32	24	137	48	29

contains 752 and 544 sentences from male and female speakers with disfluencies. The training database with filled pauses was used to train the filled pause GMMs for "ah," "ung," "um," "em," and "hem." The testing database was collected from conversational speech and consists of 756 fluent sentences and 108 sentences with filled pauses.

5.2. *Experiments on Feature Selection*

This experiment was conducted to select efficient discriminant features from the 48 analyzed features, which include 12 MFCCs, 12 delta MFCCs, 12 LPCs, F1, F2, F3, $MD(F1, Z1)$, $MD(F2, Z2)$, $MD(F3, Z3)$, FMR (F2, F1), FMR(F3, F1), FR(F3, F2), SE(F1), SE(F2), and SE(F3). In KLT analysis, all eigenvalues are calculated and the main component scree plot is shown in Fig. 10. If the value of m is chosen according to the traditional 5% residual error, it should be chosen as 29. If we use the Bartlett Chi-Square testing, as described in Section 2.4, the value of m should be chosen as 26 and the residual error becomes 8.17% under a confidence level of 97.5%. The detection rates for $m = 26$, 29, and 48 (original features) are listed in Table 4. The original 48 features show a good detection rate of 74%. This is because these features were chosen according to the special properties of filled pauses. In the performances of KLT for $m = 26$ and 29 in Table 4, $m = 26$ achieved the better result. In order to be clearer on comparing selection methods, the detection rates of KLT with different values of m were shown in Fig. 11. The detection rate decreased for $m \geq 26$ and the best performance happened when $m = 25$. Although the selection for $m = 26$ cannot achieve the best performance, this result shows that the KLT analysis with Bartlett Chi-Square testing can select the better discriminant features than the conventional method. As a result, considering the computation time and significant features, the acceptable value for m was 26 with a 76.8% detection rate.

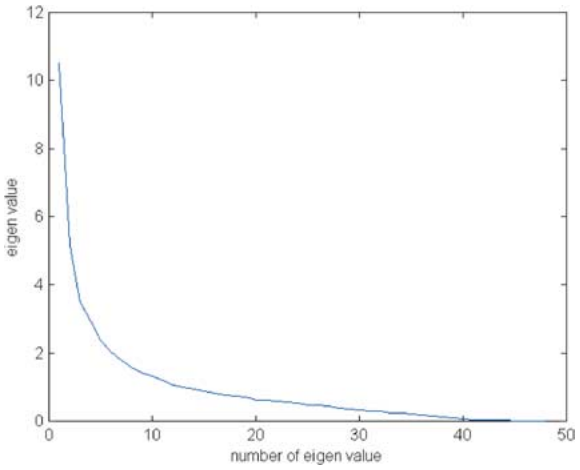


Figure 10. Main component scree plot for all eigenvalues in KLT analysis.

Table 4. Filled pause detection rates for $m = 26, 29$, and 48 without GMM.

m	48 (Original features without KLT or LDA)	29 (KLT)	26 (KLT)	26 (LDA)
Detection rate (%)	74.0	75.6	76.8	78.1

In the LDA process, the projection matrix is calculated using the $S_w^{-1}S_b$ eigenvectors. We used the 26 largest eigenvalues in KLT when the samples were well separated in the 26-dimensional space. This experimental result is also shown in Table 4. The filled pause detection rate was 78.1%. An improvement of 1.3% was obtained compared to the KLT method. It is obvious that the LDA features outperformed the KLT features in detection performance under the same feature number.

5.3. *Experiment on GMMs with KLT and LDA Features*

In order to evaluate the detection rate for GMMs with KLT and LDA features for different values of m, a frame-based filled pause detection rate was used, defined as

$$DetectionRate = 1 - (ErrorRate_{\mathrm{FA}} + ErrorRate_{\mathrm{FR}}) \tag{26}$$

where the $ErrorRate_{\mathrm{FA}}$ is the false alarm rate and $ErrorRate_{\mathrm{FR}}$ is the false rejection rate. They are defined

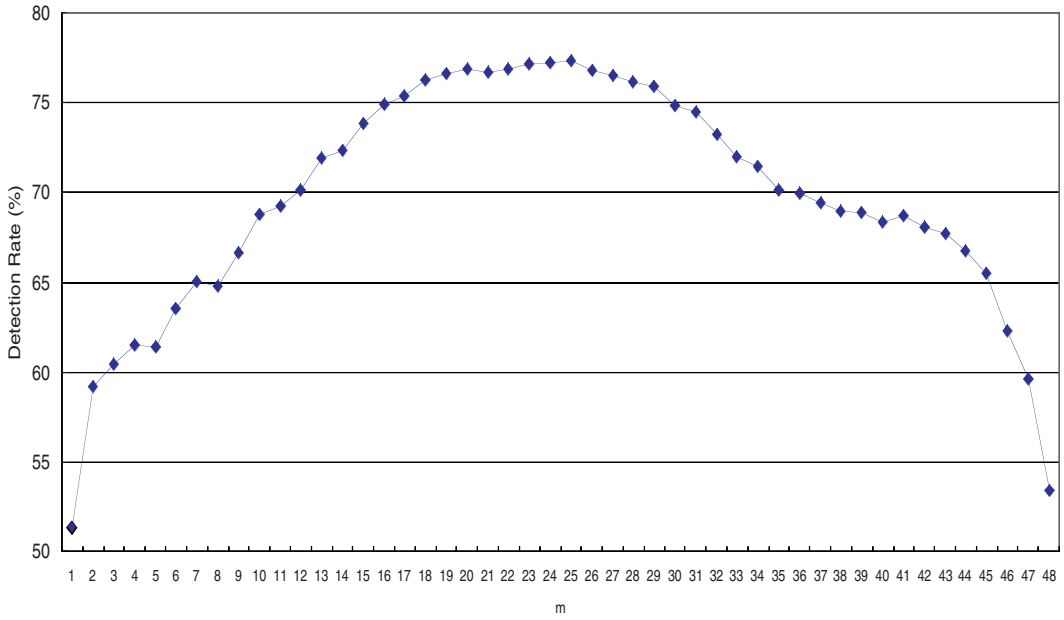


Figure 11. Detection rates of KLT with different values of m.

as follows.

$$ErrorRate_{\mathrm{FA}} = \frac{N_{\mathrm{FA}}}{N_{\mathrm{TFS}}} \tag{27}$$

$$ErrorRate_{\mathrm{FR}} = \frac{N_{\mathrm{FR}}}{N_{\mathrm{TFP}}} \tag{28}$$

where N_{TFS} is the total frame number for fluent speech patterns. N_{FA} is the frame number for fluent speech segments mis-recognized as filled pauses. Similarly, N_{TFP} is the total frame number for filled pauses, N_{FR} is the frame number of filled pauses mis-recognized as fluent speech. In the KLT analysis, the error rates for GMM with $m = 26$, 29, and 48 (original features) are shown in Fig. 12. The best detection rates for the three values of m ($m = 26$, 29, and 48) were 78.9%, 77.5%, and 75.7%, respectively. Under the same condition, the best detection rate of 79.8% for the LDA features with $m = 26$ is also shown in Fig. 12(d). This shows an improvement of 0.9% over the KLT features in detection rate and an improvement of 4.1% over the original features. It is obvious that the KLT and LDA features with $m = 26$ are more reliable for feature vector with lower dimensionality and demonstrate better performance in the experiment. The performance improvement is because: (1) KLT and LDA do serve as a decorreclation process and diagonal covariance matrices for the GMMs are used instead of the full covariance matrices. (2) With the same amount of training data, the estimates of model parameters are more reliable for feature vector with lower dimensionality since the number of these model parameters are less.

5.4. *Experiment on Discriminative GMM with KLT and LDA Features*

This experiment was conducted to show the effectiveness of discriminative training and compare the detection rates of GMM and discriminative GMM with KLT and LDA features for $m = 26$ when a threshold T is chosen. Before discriminative training, an experiment on detection rate using verification score (Eq. (21)) was conducted. The result is shown in Fig. 13. The best performances for the KLT and LDA features were 81.6% and 83.2% respectively. The verification score does improve the detection performance, but this is still not the optimal result. After discriminative training, Fig. 14 shows the detection rates for discriminative GMMs. Using

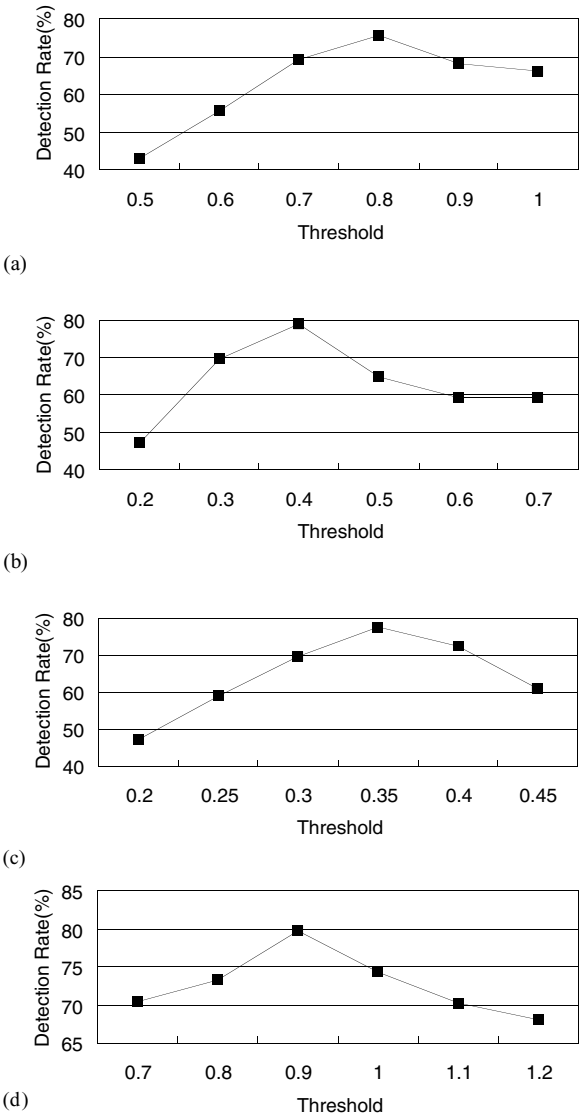


Figure 12. Detection rates using GMMs with (a) original features $m = 48$, (b) $m = 26$ (KLT), (c) $m = 29$ (KLT) and (d) $m = 26$ (LDA).

KLT features, when the threshold is $T = 2$, the maximum detection rate achieved 84.4%. The performance was further improved by 2.8%. For the LDA features, the maximum detection rate was 86.8% and the performance was further improved by 3.6%. The KLT features achieved an improvement of 5.5% and the LDA features achieved 7% improvement when discriminative training was applied to optimize the weights compared to the GMM without discriminative training.

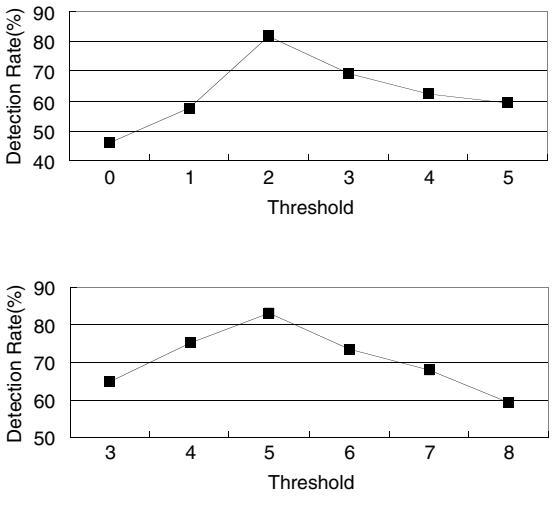


Figure 13. Detection rates using verification scores generated from GMMs with (a) KLT features ($m = 26$) and (b) LDA features ($m = 26$).

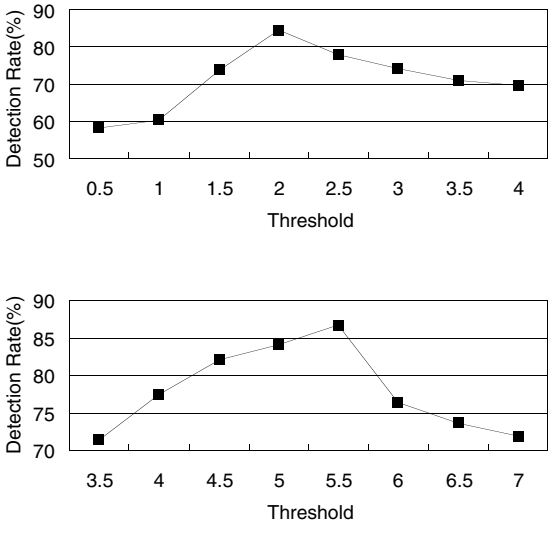


Figure 14. Detection rates using discriminative GMMs with (a) KLT features and (b) LDA features.

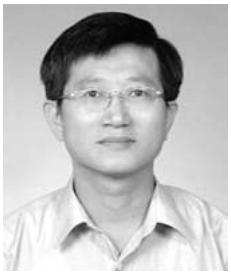
6. Conclusion

In this paper, the properties of the filled pauses "ah," "ung," "um," "em," and "hem" were analyzed. Forty-eight features that describe the properties of filled pauses were first extracted and analyzed using the KLT and LDA techniques. These analyses generated

the most discriminant features from 48 features automatically using the associated optimal linear projection. Finally the twenty-six discriminant features, called KLT and LDA features, were selected according to the Bartlett hypothesis testing. The KLT and LDA features were modeled using the GMM and a discriminative training methodology was employed to train the weights using a gradient-based iterative procedure. The experimental results show that the discriminative GMM achieved 84.4% and 86.8% detection rates for the KLT and LDA features, respectively. The results also show that a significant detection rate improvement was achieved using the discriminative GMM with KLT and LDA features.

References

1. W. Ward, "Understanding Spontaneous Speech: The Phoenix System," *Proc. of ICASSP-91*, 1991, pp. 365–367.
2. A. Kai and S. Nakagawa, "Investigation on Unknown Word Processing and Strategies for Spontaneous Speech Understanding," *Proc. of Eurospeech'95*, 1995, pp. 2095–2098.
3. A. Stolcke and E. Shriberg, "Statistical Language Model for Speech Disfluencies," *Proc. of ICASSP-96*, vol. 1, 1996, pp. 405–408.
4. M. Siu and M. Ostendorf, "Modeling Disfluencies in Conversation Speech," *Proc. of ICSLP-96*, vol. 1, 1996, pp. 386–389.
5. M. Siu and M. Ostendorf, "Variable N-Grams and Extensions for Conversational Speech Language Modeling," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 1, 2000, pp. 63–75.
6. L.M. Tomokiyo, "Linguistic Properties of Non-Native Speech," *Proc. of ICASSP-2000*, vol. 3, 2000, pp. 1335–1338.
7. M. Swerts, A. Wichmann, and R.J. Beun, "Filled Pauses as Markers of Discourse Structure," *Proc. ICSLP-96*, vol. 2, 1996, pp. 1033–1036.
8. D. O'Shaughnessy, "Recognition of Hesitations in Spontaneous Speech," *Proc. of ICASSP-92*, vol. 1, 1992, pp. 521–524.
9. M. Gabrea and D. O'Shaughnessy, "Detection of Filled Pauses in Spontaneous Conversation Speech," *Proc. of ICSLP-2000*, 2000.
10. G. Feng and E. Castelli, "Some Acoustic Feature of Nasal and Nasalized Vowels: A Target for Vowel Nasalization," *J. Acoust. Soc. Am.*, vol. 99, no. 6, 1996, pp. 3694–3706.
11. M.Y. Chen, "Acoustic Correlates of English and French Nasalized Vowels," *J. Acoust. Soc. Am.*, vol. 102, no. 4, 1997, pp. 2360–2370.
12. O. Fujimura, "Analysis of Nasal Consonants," *J. Acoust. Soc. Am.*, vol. 34, 1962, pp. 1865–1875.
13. D. Recasens, "Place Cues for Nasal Consonants with Special Reference to Catalan," *J. Acoust. Soc. Am.*, vol. 73, no. 4, 1983, pp. 1346–1353.
14. C.-H. Wu and G.-L. Yan, "Discriminative Disfluency Modeling for Spontaneous Speech Recognition," *EuroSpeech*, vol. 3, 2001, pp. 1955–1958.
15. W.R. Dillon and M. Goldstein, *Multivariate Analysis*, New York, U.S.A.: Wiley, 1984, pp. 44–46.
16. F. Beaufays, M. Weintraub, and K. Yochai, "Discriminative Mixture Weight Estimation for Large Gaussian Mixture Models," *Proc. of Acoustics, Speech, and Signal Processing*, vol. 1, 1999, pp. 337–340.
17. S. Ghaemmaghami, M. Deriche, and B. Boashash, "Hierarchical Approach to Formant Detection and Tracking Through Instantaneous Frequency Estimation," *Electronics Letters*, vol. 33, no. 1, 1997, pp. 17–18.
18. L.D. Swets and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, 1996.
19. A.M. Martinez and A.C. Kak, "PCA verus LDA," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, 2001.
20. L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, New Jersey, U.S.A.: Prentice Hall, Englewood Cliffs, 1993, pp. 271–274.



Chung-Hsien Wu received the B.S. degree in electronics engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1981, and the M.S. and Ph.D. degrees in electrical engineering from National Cheng Kung University, Tainan, Taiwan, R.O.C., in 1987 and 1991, respectively. Since August 1991, he has been with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan. He became a professor in August 1997. From 1999 to 2002, he served as the Chairman of the Department. He also worked at Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, in summer 2003 as a visiting scientist.

His research interests include speech recognition, text-to-speech, multimedia information retrieval, spoken language processing and sign language processing for hearing-impaired. Dr. Wu is a senior member of IEEE and a member of International speech communication association (ISCA) and ROCLING.
chwu@csie.ncku.edu.tw

Gwo-Lang Yan received the B.S. degree in information computer engineering from Chung-Yuan Christian University, Chung-Li, Taur

Yuan, Taiwan, in 1995, and the M.S. degree in computer science information engineering from National Cheng Kung University in 1997. He is currently pursuing the Ph.D. degreed in the Department of computer science and information engineering, National Cheng Kung University, Tainan, Taiwan, R.O.C., and the instructor at the department of information management, Kao-Yuan Institute of Technology, Kaohsiung, Taiwan. His research interests include digital signal processing, speech recognition, keyword spotting, and natural language processing.
yangl@csie.ncku.edu.tw