1355. Optical music recognition of the singer using formant frequency estimation of vocal fold vibration and lip motion with interpolated GMM classifiers

Ing-Jr Ding[1], **Chih-Ta Yen**[2], **Che-Wei Chang**[3], **He-Zhong Lin**[4]
Department of Electrical Engineering, National Formosa University, Yunlin, Taiwan
[2]Corresponding author
E-mail: [1]*ingjr@nfu.edu.tw*, [2]*chihtayen@gmail.com*, [3]*10165107@gm.nfu.edu.tw*, [4]*hajo0819@yahoo.com.tw*

(Received 18 October 2013; received in revised form 21 December 2013; accepted 28 December 2013)

Abstract. The main work of this paper is to identify the musical genres of the singer by performing the optical detection of lip motion. Recently, optical music recognition has attracted much attention. Optical music recognition in this study is a type of automatic techniques in information engineering, which can be used to determine the musical style of the singer. This paper proposes a method for optical music recognition where acoustic formant analysis of both vocal fold vibration and lip motion are employed with interpolated Gaussian mixture model (GMM) estimation to perform musical genre classification of the singer. The developed approach for such classification application is called GMM-Formant. Since humming and voiced speech sounds cause periodic vibrations of the vocal folds and then the corresponding motion of the lip, the proposed GMM-Formant firstly operates to acquire the required formant information. Formant information is important acoustic feature data for recognition classification. The proposed GMM-Formant method then uses linear interpolation for combining GMM likelihood estimates and formant evaluation results appropriately. GMM-Formant will effectively adjust the estimated formant feature evaluation outcomes by referring to certain degree of the likelihood score derived from GMM calculations. The superiority and effectiveness of presented GMM-Formant are demonstrated by a series of experiments on musical genre classification of the singer.

Keywords: musical genre classification, acoustic formant, vocal fold vibration, lip motion, Gaussian mixture model.

1. Introduction

Speech signal processing techniques, such as speech recognition [1], speaker identification [2], speaker verification [3], speech synthesizing [4], speech coding [5], and optical musical recognition (OMR) have been popular and widely-seen in lots of electronic application products. Most of those techniques focus on the data processing of acoustic vibration signals. The acoustic analysis of human vocal vibration signals is one of the most complex and important issues during the speech processing. The detection, quantification, analysis and classification of vocal fold vibrations and the lip motion have attracted much attention in medical voice assessment [6], speech communication systems [7], computer visualization [8] and optical musical recognition in this study. Figure 1 depicts the structure of human vocal fold and uttered voice signals through oscillation of human vocal fold. It is shown in Fig. 1, oscillations of vocal folds within the larynx cause human voice signals. In fact, the voice signal is constructed by the excitation of the air stream. The vibration of vocal folds provides essential information associated with phonation and has a direct impact on the properties of voices including the features of pitch, energy and formants. In the general speech pattern recognition applications including musical genre classification of the singer in this work, features of oscillated speech signals can be used for decision evaluation of recognition outcomes. The quality of the acoustic feature extracted from the vibrated speech signals will decide directly the classification performance.

This work performs musical genre classification of the singer in the area of optical music recognition using acoustic formant analysis of vibrated speech signals. Since optical musical recognition can automatically determine the musical style of the singer by optically detecting vibrated voice signals and then calculating the required formant features without any professional musicians, OMR has been an important topic in the fields of Opto-Mechatronics and pattern recognition [9-12].

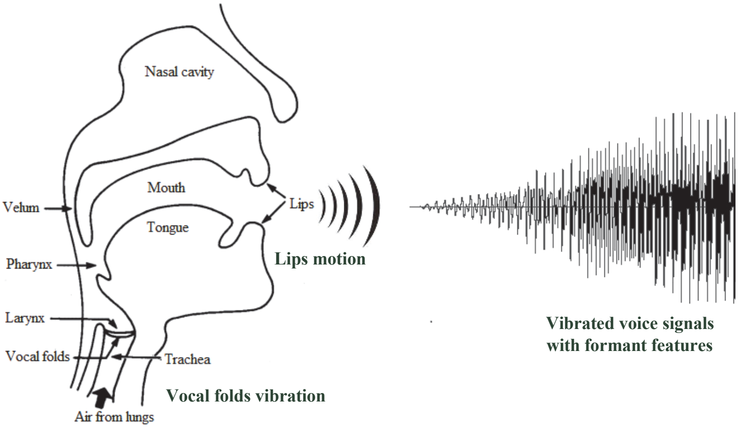


Fig. 1. The structure of vocal folds and lips of a singer, and vibrated voice signals through both vocal folds oscillation and human lips motion

This paper focuses on the areas of musical genre categorization of the singer. In fact, musical genre classification has been seen in lots of studies in the recent years. Musical genre classification applications could be viewed as a branch in the field of speech signal processing. Research pertaining to speech signals processing encompasses myriad branches including encoding/decoding, identification/verification and analysis/synthesis [13]. Musical genre classification of OMR is also frequently categorized into the class of identification/verification.

Studies on musical genre classification of the singer focus on two main classification techniques, which are feature-based and model-based categories of methods. Different to those conventional methods to use only the unique feature-based or model-based method on singer classification, this study adopts a combined feature-based and model-based method. In the feature-based category of techniques, formant frequency estimate of speech signals [14] is employed in this paper. Formant-based musical genre classification of the singer is getting more and more attention.

The category of model-based techniques for musical genre classification of the singer is to establish a statistical model and then the trained model is used to classify the input test music and decide the tendency characteristics of the music. Gaussian mixture model (GMM) [15] has been a popular classification model in the field of popular music\song classification for its excellent recognition accuracy. In fact, GMM is frequently used in the field of speaker recognition [16, 17]. The GMM classifier is a typical classification scheme of pattern recognition applications. The architecture of a typical identification system is associated with established GMM classification models, where the input test acoustic samples are segmented into the frame sequence, and from which acoustic features are extracted to estimate the likelihood degree of GMM classification models via the classifier operation. When collecting the likelihood degree estimates at a predefined time period, the classification operation is completed and the decision of the classification tendency of this input test acoustic sample can then be made. Due to the effectiveness of GMM on voice signals classification, this paper employs GMM for the model-based technique.

Although the above-mentioned feature-based and model-based classification techniques could perform well in a general music genre classification application, the classification accuracy of the system will be doubtful when encountering an adverse environment where the input test acoustic\musical data is substandard or scarce. Few studies focus on the fusion mechanism of feature-based and model-based classification techniques. In this paper, to increase the recognition accuracy of musical style classification of the singer, a fusion scheme that combines GMM model-based calculations and acoustic formant feature-based analysis, called GMM-Formant, is developed. The proposed GMM-Formant method uses acoustic formant analysis of oscillated speech signals of vocal fold vibrations and lip motions with interpolated GMM estimates to make a classification decision. GMM-Formant takes use of linear interpolation for combining GMM likelihood estimates and acoustic formant evaluation results appropriately, which will be introduced in the following sections. In this work of music genre classification of the singer, four musical styles are set, which are ‘Brisk’, ‘Rock’, ‘Lyrics’ and ‘Happy’ categories. Therefore, all the singers will be divided into four different categories, and the singer will be recognized as one of four classes.

The linear interpolation technique utilized in this work is particularly useful to deal with the problem of improper training data for GMM vocal model establishments or the problem of scarce test data for musical style recognition of the GMM classifier. Many approaches for speaker adaptation in the field of automatic speech recognition also employ such linear interpolation techniques, such as MAP of the Bayesian-based adaptation category, MAPLR of the transformation-based adaptation category and eigenspace-based MLLR (eigen-MLLR) of Eigenvoice adaptation category [18]. MAP estimate is a typical representative of Bayesian-based adaptation that uses the linear interpolation framework [19]. Following the idea of MAP estimate, the developed GMM-Formant for optical musical recognition utilizes the similar formulation of linear interpolation where the amount of test data available for recognition is specifically taken into account in linear interpolation design.

2. Feature-based and model-based singer classification

Generally, automatic musical style recognition of the singer in OMR could be performed by two main categories of techniques, feature-based and model-based approaches. In feature-based music recognition, the formant parameter obtained from vocal fold vibrations and lip motions is an important acoustic feature for evaluation. In model-based classification techniques, GMM is widely used for its simplicity. Two main classification techniques, model-based and feature-based determination mechanisms, for singer classification recognition are adopted in this work for their effectiveness and efficiency on classification performances. This section will describe the widely-used formant feature analysis of vibrated speech signals and the popular Gaussian mixture model classification methods.

Formants are essentially the free resonance of the human vocal-tract system [14]. As mentioned before, voice signals are made through oscillation of human vocal fold and formant is one of the most important acoustic features. Formant is the regional frequency of the sound energy and can calculate the low frequency region that the person’s ear could hear. Formant puts extremely little emphasis on a high and rough frequency region. Formant spectrum was produced by calculating the input musical data. Formant spectrum contains lots of peak values in every spectrum. For example, there are totally N formant in the spectrum, generally denoting as F_1, F_2, F_3,…, F_N, each of which represents a different frequency and energy.

Linear predictive coding (LPC) is the most popular formant estimation technique. LPC is an important procedure during feature extraction of traditional speech recognition. Assume that $x(n)$ is the original speech signals, the LPC form of $x(n)$ is shown as follows:

$$LPC\ form\ of\ speech\ signals = \sum_{i=1}^{P} a_i \cdot x(n-i), \tag{1}$$

where a_i, $i = 1,2,\dots,P$, is the LPC coefficients. For obtaining a_i an autocorrelation approach is adopted where the error between the original speech signals $x(n)$ and the LPC form of $x(n)$ is to be minimized, and the following relation is constructed:

$$r_x(\tau) = \sum_{i=1}^{P} a_i \cdot r_x(\tau - i), \ \ \tau = 1,2,\dots,L, \tag{2}$$

where $r_x(\tau)$ is the autocorrelation function of the original speech signals $x(n)$ with lag values $\tau = 1,2,\dots,L$. Given the autocorrelation function as shown in Eq. (2), the coefficient parameters a_i can be determined by solving a set of linear equations as follows:

$$\begin{bmatrix} r_x(0) & r_x(-1) & \cdots & r_x(-P+1) \\ r_x(1) & r_x(0) & \cdots & r_x(-P+2) \\ \vdots & \vdots & & \vdots \\ r_x(P-1) & r_x(P-2) & \cdots & r_x(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} r_x(1) \\ r_x(2) \\ \vdots \\ r_x(P) \end{bmatrix}. \tag{3}$$

In formant feature analysis, an all-pole mode for defining the characteristics of a song before calculating the envelopment is necessary, which is shown as follows [14]:

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i \cdot z^{-i}}. \tag{4}$$

In Eq. (4), the parameter G, called as gain factor, is a constant and usually set as 1, the parameter P denotes the linear predictive order and a_i means the linear prediction coefficients, also known as the autoregressive parameters of the filter $A(z)$, which could be derived from Eq. (3). For simplicity on practical evaluation of input acoustic signals, the above formant feature analysis is transformed into the following operation formula:

$$frequency(x) = \frac{f_s}{framesize} \cdot index, \tag{5}$$

where $frequency(x)$ is the frequency of the xth peak value, f_s denotes the sampling rate, $framesize$ means the length of a frame and the parameter $index$ is the index value of a peak value.

The framework of a popular song classification system that uses the model-based technique is associated with established audio models, frequently-seen GMM for example, where the input popular music from the database is segmented into the frame sequence, and from which audio features are extracted to evaluate the characteristic and the classification tendency of this music via classification operations cooperated with GMM audio model calculations.

Mathematically, a GMM is a weighted sum of M Gaussians, denoted as [15]:

$$\lambda = \{w_i, \mu_i, \Sigma_i\}, \ \ i = 1, 2, \dots, M, \ \ \sum_{i=1}^{M} w_i = 1, \tag{6}$$

where w_i is the weight, μ_i is the mean and Σ_i is the covariance. In this study, four parameter sets, λ_1, λ_2, λ_3 and λ_4 (four audio GMM models, that is), for representing the musical characteristics of four different categories, ‘Brisk’, ‘Rock’, ‘Lyrics’ and ‘Happy,’ are determined, respectively.

After completing the training of the four audio GMM models, the recognition procedure can then be executed based on these four trained GMM models. Note that the musical genre classifier deployed here is a GMM classifier consisting of four separate audio GMM models: the first is for the ‘Brisk’ classification, the second is for the ‘Rock’ classification, the third is for the ‘Lyrics’ classification and the last is for the ‘Happy’ classification. Consider the classifier operating with a decision window (or equivalently, over a time interval) covering n audio feature vectors of D dimensions, $X = \{x_i | i = 1,2, \dots, n\}$, combined with four audio classification models, λ_1, λ_2, λ_3 and λ_4.

During the recognition phase of musical style classification, the class of X is determined by maximizing a posteriori probability $P(\lambda_s | X)$:

$$\hat{s} = \underset{s=\{1,2,3,4\}}{\operatorname{argmax}} P(\lambda_s | X) = \underset{s=\{1,2,3,4\}}{\operatorname{argmax}} \frac{L(X|\lambda_s)}{L(X)} \cdot P(\lambda_s). \tag{7}$$

Note that:

$$L(x_i|\lambda_s) = \sum_{j=1}^{M} w_j \cdot \frac{1}{(2\pi)^{D/2} \cdot \left|\Sigma_{s_j}\right|^{1/2}} \cdot \exp\left\{-\frac{1}{2}\left(x_i - \mu_{s_j}\right)^T \left(\Sigma_{s_j}\right)^{-1} \left(x_i - \mu_{s_j}\right)\right\}. \tag{8}$$

In real implementation, Eq. (7) is replaced by:

$$\hat{s} = \underset{s=\{1,2,3,4\}}{\operatorname{argmax}} \sum_{i=1}^{n} \log L(x_i|\lambda_s), \tag{9}$$

for simplicity. In addition, at the end of the recognition procedure, the test audio signal X is then classified as one of four musical categories indicated by $\hat{s}$.

3. Proposed GMM-formant using acoustic formant features with interpolated GMM estimation for singer classification

Following the thought line of MAP estimate performed on speech recognition, the proposed GMM-Formant method for optical musical recognition employs MAP-like linear interpolation where the amount of test data available will be specifically taken into account in linear interpolation design.

MAP estimate is a type of direct speech recognition model adjustments that attempts to re-estimate the model parameters directly [19]. MAP offers a framework of incorporating newly acquired speaker-specific data into the existing models. However, MAP re-estimates only a portion of the model parameter units associated with the adaptation data. Therefore, MAP estimate usually requires a significant amount of data. The recognition performance of speech recognition is improved as the adaptation data increase and the adaptation gets covering the model space. When a sufficient amount of data is available, the MAP estimation yields recognition performances equal to those obtained using maximum-likelihood estimation. The linear interpolation form of MAP is shown in Eq. (10):

$$\hat{\mu}_k = w_1 \cdot \bar{y}_k + w_2 \cdot \mu_k, \tag{10}$$

where:

$$w_1 = \frac{N_k}{\tau + N_k}, \tag{11}$$

$$w_2 = \frac{\tau}{\tau + N_k}, \tag{12}$$

$$w_1 + w_2 = 1. \tag{13}$$

Observed from Eq. (10) to Eq. (13), the MAP estimate of the mean is essentially a weighted average of the prior mean (μ_k) and the sample mean ($\bar{y}_k$), and the weights are functions of the number of adaptation samples (N_k) if τ is fixed. When N_k is equal to zero (i.e., no additional training data are available for adapting the kth Gaussian), the estimate is simply the prior mean of the kth Gaussian alone. Conversely, when a large number of training samples are used for the kth Gaussian ($N_k \rightarrow \infty$, to be exaggerative), the MAP estimate in Eq. (10) then converges asymptotically to the maximum likelihood estimate, i.e., the sample mean parameter with the kth Gaussian, $\bar{y}_k$.

The proposed GMM-Formant for musical genre classification employs the MAP-like interpolation. As mentioned in the previous section, the operation procedure of GMM classifier performs a fast recognition classification calculation using the simple likelihood calculation as shown in Eq. (9) to obtain the likelihood score between the input acoustic song data and the song classification model. When given proper song data with standard property, classification operation by using the GMM classifier is effective. However, given substandard song data for recognition classification, the accuracy of the estimated likelihood score using Eq. (9) is dubious. Poor estimation of the likelihood score in turn leads to incorrect recognition results on song classification. The problem of improper testing data for GMM classifier classification operation can be alleviated by additionally referring to certain degree of evaluation information of feature-based classification operation using a MAP-like linear interpolation scheme.

Given substandard test song data for classification, it is necessary to be more "conservative" in using the derived likelihood score for classification decisions. In other words, the effect of the improper data should be restricted so that the final decision does not reference too much from the model-based classification calculation outcome derived by GMM classifier estimate. Therefore, this study proposes the GMM-Formant approach which combines both model-based and feature-based classification operations as follow:

$$Adjusted\ Likelihood\ Score = \alpha \cdot GMM + (1 - \alpha) \cdot Formant, \quad 0 \leq \alpha \leq 1, \tag{14}$$

where GMM is the likelihood score calculated by GMM classifier operation, i.e., Eq. (9) in the previous section and $Formant$ is the estimated formant frequency information derived from formant feature analysis as described in the previous section. The likelihood score for song classification decisions is not determined by only the GMM likelihood estimate. Instead, this proposed approach as shown in Eq. (14) calculates a weighted sum of the GMM model-based estimate and the formant feature-based evaluation. The form of linear interpolation in Eq. (14) is used to tune the GMM likelihood score derived from Eq. (9), and with the proper adjustment of formant evaluated information, the final adjusted GMM likelihood score for recognition classification decision on input test songs will be more reliable and believable.

Observed from the designed interpolation formula in Eq. (14), the interpolation form of proposed GMM-Formant behaves as that of the above-mentioned MAP estimate. A weight parameter α governs the balance of GMM and $Formant$, mimicking the role of the parameter α for tuning the likelihood score from the GMM classifier. Using a weighting scheme with the adjustable parameter α should achieve satisfactory recognition classification performance even when encountering improper test song data for classification decisions. Note that the weight α varies depending on how much confidence one has in the likelihood score derived from GMM classifier estimate. A possibly not so well estimate of the likelihood score calculated from Eq. (9) due to substandard test song data would preferably goes with α approaching 0 so that the biased estimate of the GMM likelihood estimate will be restricted. Conversely, 1-approaching α should be given.

As the suggestion of MAP estimate in Eq. (10) to Eq. (13), the weight parameter α in Eq. (14) could be further designed as follows:

$$\alpha = \frac{N}{\sigma + N}, \tag{15}$$

where N is the data size of the test audio data for musical genre recognition (i.e., the total number of the audio frames) and σ denotes the weight control parameter. The parameter σ in Eq. (15) can be used to control the balance between GMM and formant recognition evaluation outcomes. When N is large (i.e., large-sized test data are available for musical style recognition), the GMM evaluation result that accumulates more likelihood estimates is therefore more reliable, the adjusted likelihood score is simply the GMM classification evaluation outcome alone. Conversely, when only a small number of test samples for musical genre recognition are available, developed GMM-Formant then approaches to the side of formant evaluation results. Now consider the other way round with N being fixed, the parameter σ controls the balance in the interpolation between the GMM-term and the $Formant$-term, (as N does). It could be viewed as the weight control parameter in that the balance between the GMM-term and the $Formant$-term can be achieved by choosing a proper value of σ. The parameter σ determines, to which side of, and for how close to the GMM-term or the $Formant$-term, the $Adjusted\ Likelihood\ Score$-term for recognition decision would be.

4. Experiments and results

The experiments on musical genre classification of the singers with the proposed GMM-Formant method are performed in a database which contains 50 Mandarin popular songs. These 50 Mandarin popular songs are recorded by a group of the designed singers.

The analysis frames were 20-ms wide with a 10-ms overlap. For each collected song with PCM form, the wave header is then added to the front side of the PCM raw data. The related settings of each song with wave form were 1411 kbps (bitrates), 16 bits (resolutions), stereo (channels) and 44100 samples per second (sampling rate). The analysis frames were 20-ms wide with a 10-ms overlap. For each frame, a 10-dimensional feature vector was extracted. The feature vector for each frame was a 10-dimensional cepstral vector.

Table 1. The arranged group of singers with four categories of lip motions, each of which is to generate one of 'Brisk', 'Rock', 'Lyrics' and 'Happy' songs in the training phase

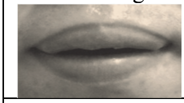
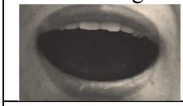
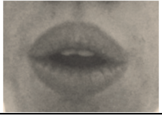
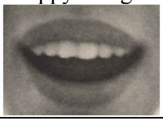
Musical genre	Songs that collected by the designed group of singers (in Mandarin)
'Brisk' Singers 	"哇哈哈,""牛仔很忙,""撐腰,""夏日瘋,""向前衝,""當我們宅一塊,""離開地球表面,""轟炸,""稻香,""高高在下"
'Rock' Singers 	"音浪,""王妃,""鬧翻天,""我秀故我在,""全面失控,""Wake Up,""超級右腦,""強心臟,""舞極限,""Count on Me"
'Lyrics' Singers 	"Baby Tonight,""王寶釧苦守寒窯十八年,""凌晨三點鐘,""末班車,""累,""忘記擁抱,""爸爸媽媽,""城裡的月光,""最寂寞的時候,""還是愛著你,"
'Happy' Singers 	"You Are My Baby,""快樂頌,""歐拉拉呼呼,""歐兜拜,""春天的吶喊,""太空警察,""就像白癡一樣,""完美男人,""麻吉麻吉,""要去高雄"

Table 2. Musical content in the testing phase for singer classification

Musical genre	Titles of popular songs (in Mandarin)
Brisk, Rock, Lyrics or Happy	"愛走秀,""億萬分之一的機率,""慢靈魂,""Super Nice Girl,""3D 舞力全失,""星晴,""驚嘆號,""超跑女神,""漂流瓶,""太熱"

Table 3. Recognition accuracy of proposed GMM-Formant with various values of α in the testing experiment of singer classification

Settings of α value	Recognition rate (%)
$\alpha = 0.1$	45
$\alpha = 0.2$	52.5
$\alpha = 0.3$	52.5
$\alpha = 0.4$	52.5
$\alpha = 0.5$	52.5
$\alpha = 0.6$	62.5
$\alpha = 0.7$	65
$\alpha = 0.8$	67.5
$\alpha = 0.9$	72.5

Table 4. Recognition performance comparisons of conventional GMM, conventional acoustic formant analysis, proposed GMM-Formant in the testing experiment of singer classification

Settings of α value	Recognition rate (%)
Conventional GMM	65
Acoustic formant analysis	60
GMM-Formant with $\alpha = 0.9$	72.5

The database was composed of four types, ‘Brisk’, ‘Rock’, ‘Lyrics’ and ‘Happy’ singers. Each song recorded by the singer in the database belongs one of these four genres. The experiments of musical genre classification of the singers were divided into two phases, the training phase and the testing phase. Table 1 and Table 2 show the musical contents of the training data and the testing data respectively. Each popular Mandarin song collected by one designed singer with the specific musical style of singing was categorized to one of ‘Brisk’, ‘Rock’, ‘Lyrics’ and ‘Happy’ musical genres. As could be seen in Table 1, the singing persons that are classified as one of four categories, ‘Brisk’, ‘Rock’, ‘Lyrics’ and ‘Happy’ classes, have different lips motions.

The performances of the proposed GMM-Formant with various values of α on singers classification are shown in Table 3. Observed from Table 3, proposed GMM-Formant with the setting of $\alpha = 0.9$ has the highest recognition rate, which achieves to 72.5 %. Conversely, an improper setting of α will restrict the GMM-Formant method. For example, when $\alpha = 0.1$ is set to GMM-Formant, the recognition accuracy is dissatisfactory, which is 45 % and uncompetitive. Experimental results from Table 3 reveal that proposed GMM-Formant has a better and more acceptable performance on recognition accuracy when the value of α is set higher. As mentioned in the previous section, the weight parameter α governs the balance of two terms GMM and $Formant$, and Table 3 suggests that an 1-approaching α set for GMM-Formant to form a weighted sum of the GMM model-based estimate and the formant feature-based evaluation will be a positive tendency.

The competitiveness of the proposed GMM-Formant method is demonstrated in Table 4. From Table 4, proposed GMM-Formant with the setting of $\alpha = 0.9$ performs best. Proposed GMM-Formant is better than conventional GMM and acoustic formant analysis on recognition performances by 7.5 % and 12.5 %, respectively.

5. Conclusions

This paper proposes a GMM-Formant scheme for performing music genre classification of the singer. Optical musical recognition of the singer with proposed GMM-Formant will be as a more intelligent analytical tool for automatic classification of the singers. The proposed GMM-Formant takes use of the popular linear interpolation technique to perform a proper fusion between model-based and feature-based classification processing. In developed GMM-Formant, Gaussian mixture model is adopted for model-based classification, and acoustic formant feature analysis is utilized to carry out feature-based classification. The presented GMM-Formant effectively overcomes the problem of the weakness of the unique classification technique. Experimental results demonstrated that developed GMM-Formant achieved competitive and acceptable performances on classification accuracy of singers.

Acknowledgements

This research is partially supported by the National Science Council (NSC) in Taiwan under Grant NSC 101-2815-C-150-017-E. The author also gratefully acknowledges the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] **Gaikwad S. K., Gawali B. W., Yannawar P.** A review on speech recognition technique. International Journal of Computer Applications, Vol. 10, Issue 3, 2010, p. 16-24.

[2] **Jung J., Kim K., Kim M. Y.** Advanced missing feature theory with fast score calculation for noise robust speaker identification. Electronics Letters, Vol. 46, Issue 14, 2010, p. 1027-1029.

[3] **Kim K., Kim M. Y.** Robust speaker recognition against background noise in an enhanced multi-condition domain. IEEE Transactions on Consumer Electronics, Vol. 56, Issue 3, 2010, p. 1684-1688.

[4] **Raitio T., Suni A., Yamagishi J., Pulakka H., Nurminen J., Vainio M., Alku P.** HMM-based speech synthesis utilizing glottal inverse filtering. IEEE Transactions on Audio, Speech and Language Processing, Vol. 19, Issue 1, 2011, p. 153-165.

[5] **Ramamurthy K. N., Thiagarajan J. J., Spanias A.** An interactive speech coding tool using LabVIEW™. Proc. IEEE Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop, 2011, p. 180-185.

[6] **Lin C. S., Chang S. F., Chang C. C., Lin C. C.** Microwave human vocal vibration signal detection based on Doppler radar technology. IEEE Transactions on Microwave Theory and Techniques, Vol. 58, Issue 8, 2010, p. 2299-2306.

[7] **Falk T. H., Chan J., Duez P., Teachman G., Chau T.** Augmentative communication based on realtime vocal cord vibration detection. IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 18, Issue 2, 2010, p. 159-163.

[8] **Lohscheller J., Eysholdt U., Toy H., Döllinger M.** Phonovibrography: mapping high-speed movies of vocal fold vibrations into 2-D diagrams for visualizing and analyzing the underlying laryngeal dynamics. IEEE Transactions on Medical Imaging, Vol. 27, Issue 3, 2008, p. 300-309.

[9] **Rebelo A., Fujinaga I., Paszkiewicz F., Marcal A. R. S., Guedes C., Cardoso J. S.** Optical music recognition: state-of-the-art and open issues. International Journal of Multimedia Information Retrieval, Vol. 1, Issue 3, 2012, p. 173-190.

[10] **Bellini P., Bruno I., Nesi P.** Assessing optical music recognition tools. Computer Music Journal, Vol. 31, Issue 1, 2007, p. 68-93.

[11] **Bainbridge D., Bell T.** The challenge of optical music recognition. Computers and the Humanities, Vol. 35, Issue 2, 2001, p. 95-121.

[12] **Byrd D., Schindele M.** Prospects for improving OMR with multiple recognizers. Proc. International Conference on Music Information Retrieval, 2006, p. 41-46.

[13] **Wutiwiwatchai C., Furui S.** Thai speech processing technology: A review. Speech communication, Vol. 49, Issue 1, 2007, p. 8-27.

[14] **Welling L., Ney H.** Formant estimation for speech recognition. IEEE Transactions on Speech Audio Process, Vol. 6, Issue 1, 1998, p. 36-48.

[15] **Reynolds D. A., Rose R. C.** Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing, Vol. 3, 1995, p. 72-83.

[16] **You C. H., Lee K. A., Li H.** An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition. IEEE Signal Processing Letters, Vol. 16, Issue 1, 2009, p. 49-52.

[17] **Kenny P., Boulianne G., Ouellet P., Dumouchel P.** Speaker and session variability in GMM-based speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, Issue 4, 2007, p. 1448-1460.

[18] **Shinoda K.** Acoustic model adaptation for speech recognition. IEICE Transactions on Information and Systems, Vol. E93-D, Issue 9, 2010, p. 2348-2362.

[19] **Lee C. H., Lin C. H., Juang B. H.** A study on speaker adaptation of the parameters of continuous density hidden Markov models. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 39, Issue 4, 1991, p. 806-814.

Ing-Jr Ding was born in Taipei, Taiwan, in 1975. He received the B.S. degree from Chang-Gung University in 1999, M.S. degree from National Central University in 2001, and Ph.D. degree from National Chiao-Tung University in 2008. He joined the Graduate Institute of Automation and Control at National Taiwan University of Science and Technology as a project assistant professor from March 2009 to July 2009. From August 2009 to July 2012, he served as an assistant professor in the Department of Electrical Engineering, National Formosa University. Since August 2012, he has been an associate professor in the Department of Electrical Engineering, National Formosa University. His research interests include speech processing, pattern recognition, machine learning, artificial intelligence, and multimedia techniques.

Chih-Ta Yen was born in Taipei, Taiwan, in January 1974. He received his B.S. degree from the Department of Electrical Engineering at Tamkang University, Taiwan, in 1996, his M.S. degree from the Department of Electrical Engineering, National Taiwan Ocean University, Taiwan, in 2002, and his Ph.D. degree from the Department of Electrical Engineering at National Cheng Kung University, Taiwan, in 2008. He is currently an associate professor in National Formosa University in the area of communication technology at the Department of Electrical Engineering, Yunlin, Taiwan. His major interests are in the areas of multi-user optical communications, wireless communication systems, sensing systems, and satellite communications.

Che-Wei Chang received the B.S. degree from National Formosa University in 2012. Since August 2012, he has pursued the master’s degree in the Department of Electrical Engineering, National Formosa University.

He-Zhong Lin received the B.S. degree from National Formosa University in 2013. From Jul. 2012 to Feb. 2013, he was a project research assistant in the Department of Electrical Engineering, National Formosa University.