Consolidation of Robust Speaker and Speech Recognition for Intelligent Doorway Application

Ta-Wen Kuan, Jhing-Fa Wang, Po-Yi Shih, Ta-Wei Sun and Miao-Hai Chen
Department of Electrical Engineering, National Cheng Kung University
Email:gwam.davin@gmail.com, wangjf@mail.ncku.edu.tw

Abstract

In this paper, integration of speaker identification and speech recognition for intelligent doorway application has been proposed. Two target speakers will be identified through an one-word speech utterance. Moreover, this utterance will be recognized to be a pre-defined speech command. The speaker identification in the proposed framework is based on support vector machine (SVM). The “one-versus-one” approach is applied in this paper to classify test point input utterance according to the number of votes. As for the speech recognition, we use confusion matrix to develop an efficient phonetic set for a command-based multi-lingual system, the confusion matrix calculates acoustic similarities between every two phonemes. The proposed framework has been realized in the intelligent doorway application and will be applied to many other daily life computer speech applications.

Keywords: SVM, confusion matrix, HTK, speaker identification, speech recognition.

1. Introduction

In the real world, there are three commonly applications in speech recognition system, such as “who is speaker?”, “what is content?”, and “where is speaker?”. The contribution of this paper is to propose a practical consolidated framework to integrate both the speaker identification and speech recognition, with the aim at satisfaction of human computer interface in recognizing “who is speaker?” and “what is content?” at same time.

Support Vector Machine has been explored and proved in speaker recognition for many years [1][2]. SVM has many desirable attributes that can classify and robust to sparse data without over-training and to make linear and non-linear decision via kernel functions [3]. However, due to complicated algorithm and time-consuming process in training SVM, thus it still not gained widespread utilization in many applications. Ubiquitous Robot Companion (URC) proposed a text-independent speaker identification using microphone-array on a robot and intends to enrich the interaction between human and robot [4]. Far-field speaker recognition proposed two approaches to improve the robustness of speaker recognition. The first is to use the conventional method based on acoustic feature. The second approach is to make use of higher-level linguistic feature. However, the adverse environmental condition and adverse training-testing conditions still need to be considered and conquered under proposed benchmark environment [5]. Ubiquitous and robust text-Independent speaker recognition [6] proposed a new microphone-array configuration of framework for benchmark. This framework is used a mixer to received speech signal from six microphones, then the six channel speech signal are mixed and output only one signal for feature extraction.

The mixed-language speech recognition has been researched for many years [7][8][9]. In this proposed consolidated of speech recognition system, the speaker independent voice command recognition is adopted, and with a string size of tens or more words. In addition, an acoustic and phoneme modeling based on confusion matrix for ubiquitous mixed-language speech of recognition system is integrated in proposed framework [10]. This system allows users to use given command to control electrical device via speech. The system is also flexibly applied in different command-based control applications by changing the dictionary description and grammar in each new work.

The reminder of this paper is organized as follows. In Section 2, the basic theories of SVM algorithm for data classification as well as confusion Matrix of acoustic model for bilingual speech recognition are described. In Section 3, the proposed framework of consolidated speech recognition system is presented. The experimental results of proposed architecture are shown in Section 4. Finally, we draw our conclusion in Section 5.

2. Literature Review

2.1 SVM based Speaker Identification

The main concept of SVM is to use a partition hyperplane to maximize the distance between support vectors of two classes features, and then to create a classifier between two clusters of sample. The gain of the SVM-based pattern recognition method is robust to sparse training data samples [11] [12]. This optimal hyperplane is obtained by minimizing the following constrained optimization problem as shown in Eq. (1).

$$\begin{aligned}
&\min_{w,b,\xi} \frac{1}{2} w^T w + C\left(\sum_{i=1}^{N} \xi_i \right) \\
&\textit{subject to} \qquad\qquad\qquad\qquad , \qquad\qquad (1) \\
&\qquad y_i(w\phi(x_i)+b)+\xi_i-1 \ge 0, \quad 1 \le i \le n \\
&\qquad \xi_i \ge 0, \qquad\qquad\qquad\qquad 1 \le i \le n
\end{aligned}$$

where x_i is a training sample, y_i is the corresponding target value, $w \in R_m$ is a vector of weights of training instances, b is a constant, C is a real value cost parameter, and ξ_i is a penalty parameter (slack variable).

If $\Phi(x_i) = x_i$, the SVM finds a linear separating hyperplane with the maximal margin. If Φ maps x_i into a higher dimensional space, then it is called a nonlinear SVM. For the nonlinear SVM, the dimension of the vector w can be large or even infinite.

The constrained optimization problem in Eq. (1) can be handled by Lagrange multiplier approach. The Lagrange function is constructed as Eq. (2)

$$\begin{aligned} L(w,b,\xi_i,\alpha,\mu) &= \frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i \\ &\quad -\sum_{i=1}^{N}\alpha_i[y_i(w^T x_i + b) + \xi_i - 1] - \sum_{i=1}^{N}\mu_i\xi_i \end{aligned} \tag{2}$$

where α_i, μ_i are the Lagrange multipliers.

Based on the duality theorem, Eq. (2) are the primal problem and its corresponding dual is formulated as in Eq. (3).

$$\begin{aligned} &\min_{w,b,\xi} \quad \frac{1}{2}w^{\mathrm{T}}w + C\sum_{i}{}_{i=1}^{N}\xi_i \\ &subject\,to \quad y_i(w^{\mathrm{T}}x_i - b) + \xi_i - 1 \geq 0\,,\ 1 \leq i \leq N \\ &\qquad\qquad\quad \xi_i \geq 0, \qquad\qquad 1 \leq i \leq N \end{aligned} \tag{3}$$

where $C > 0$ is the upper bound of the Lagrange multipliers, ζ is penalty, b is bias, N is number of training data $\{(x_1,y_1), (x_2,y_2), (x_3,y_3),\ldots.., (x_i,y_i)\}$, w is coefficients vector and $y \in \{\pm 1\}$.

The objective function of the dual problem in Eq.(3), can be formulated and summarized as the Eq.(4) by vanished the primal variables of w, b and ζ.

$$\begin{aligned} &\max imizeL_D \equiv \sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i x_j \\ &subject \quad to \\ &\qquad 0 \leq \alpha \leq C \\ &\qquad \sum_i \alpha_i y_i = 0 \end{aligned} \tag{4}$$

Thus the solution of objective function is given as in Eq. (5)

$$w = \sum_{i=1}^{N}\alpha_i y_i x_i \tag{5}$$

To train the SVM is to search through the feasible region of the dual problem and maximize the objective function, and the optimal solution can be checked using the KKT conditions. The further detail training algorithm is described in [13].

Alternatively, the classification approaches in SVM classifier is essential to be stressed as 1).One-versus-the rest approach. 2).One-versus-one approach [9]. The first approach is to construct K separate SVMs, in which the k^{th} model $y_k(x)$ is trained using the data from class C_k as the positive examples and the data from the remaining K-1 classes as the negative examples. The second approach is to train K(K-1)/2 different 2-class SVMs on all possible pairs of classes, and then to classify test point according to which class has the highest number of 'votes'. In this paper, the one-to-one approach is adopted in our proposed framework in testing phase for speaker identification.

2.2 Confusion Matrix of Acoustic Model for Bilingual Speech Recognition

The confusion matrix is basically a confusion matrix, which is a supervised learning skill in the field of artificial intelligence and pattern recognition [10]; the confusion matrix is also called a matching matrix as well in unsupervised learning. Each column of the confusion matrix is defined as the instances in a predicted class, while each row is defined as the instances in an actual case class. The advantage of confusion matrix is simple to be observed if the system is confusing two classes. The example of confusion matrix is shown in Table.2.1.

Table.2.1. the example of confusion matrix.

		Actual case class		
		[m]	[d]	[b]
Predicted class	ㄇ	90	5	5
	ㄉ	0	100	0
	ㄅ	10	0	90

The rows of the confusion matrix are always normalized by summarized number of total symptoms for evaluation. And the value of the confusion frequency in the estimated matrix is as the relative number of confusion. Eq.(6) is given by

$$\hat{s} = \frac{card\{k:\Omega_k^{k_i}\ \ is\ \ \ classified\ \ \ as\ \ \ k_i\}}{card\{k:\Omega_k^{k_i}\}} \tag{6}$$

where card is defined as number of elements. $\hat{s}$: is a confusion matrix estimation which is obtained for the set of models, it contains the estimation of how likely it is that a given model is classified as other model.

The procedures of the mixed-language acoustic model based on Mandarin and English are shown in follow:

1). Clustering the similarity phones set acoustically and phonetically in English and Mandarin.

2). The monophonic sets are built by single Gaussian acoustic model.

3). For each phone in Mandarin, we calculate the dissimilarity of the phone set based on the confusion matrix to all the phones in the same group for English. If the value is below a threshold, the source phone in Mandarin would be mapped to that phone in English. Otherwise, both the phones would be modeled separately in the bilingual system.

4). If some phones in Mandarin can not map to phone cluster in English, in such cases will not try to map this phone in mandarin to English.

5). While the list of phones in bilingual system is finished, the lexicon for Mandarin is edited by using the mapping rules. The mixed-language phone set is shown in the Table 2.2.

Table 2.2. The mixed-language phone set

I N X	M P A	Eng P.h.	Model	I N X	M P A	Eng P.h.	Model	I N X	M P A	Eng P.h.	Model
1	ㄅ	[b]	B	20	ㄘ		C	39		[ʤ]	JH
2	ㄆ	[p]	P	21	ㄙ	[s]	S	40		[θ]	TH
3	ㄇ	[m]	M	22	ㄚ	[ɑ]	AA	41		[h]	HH
4	ㄈ	[f]	F	23	ㄛ	[o]	OW	42		[w]	W
5	ㄉ	[d]	D	24	ㄜ	[ə]	@	43		[r]	R
6	ㄊ	[t]	T	25	ㄝ	[ɛ]	EH	44		[j]	Y
7	ㄋ	[n]	N	26	ㄞ	[aɪ]	AY	45		[ɪ]	IH
8	ㄌ	[l]	L	27	ㄟ	[e]	EY	46		[ʊ]	UH
9	ㄍ	[g]	G	28	ㄠ	[aʊ]	AW	47		[ʌ]	AH
10	ㄎ	[k]	K	29	ㄡ		OU	48		[ɝ]	ER
11	ㄏ		H	30	ㄢ		AN	49		[ɔ]	AO
12	ㄐ		J	31	ㄣ		EN	50		[æ]	AE
13	ㄑ		Q	32	ㄤ		ANG	51		[δ]	DH
14	ㄒ		X	33	ㄥ	[ŋ]	NG	52		[ʃ]	SH
15	ㄓ		Zh_m	34	ㄦ		ER	53		[ʒ]	ZH
16	ㄔ		Ch_m	35	ㄧ	[i]	IY	54		[v]	V
17	ㄕ		Sh_m	36	ㄨ	[u]	UW	55		[ɔɪ]	OY
18	ㄖ	[z]	ZR	37	ㄩ		YU				
19	ㄗ		Z	38		[ʧ]	CH				

3. Proposed Framework

3.1 Proposed Consolidation of Speech Recognition Framework

The appealing work of this paper is to propose a framework, which is integrated the speaker identification and speech recognition into a consolidated speech recognition system, this architecture is shown in Fig 3.1. This system is capable of dealing with one specified word of speech signal as prior defined, then using SVM based speaker identification procedure to find out the target speaker, at the same time the same speech signal is then used to examine second target speaker by confusion matrix based of speech identification system. The scenario is such as: The Speaker A pronounces a sentence as, "I want to leave message to Speaker B" to proposed system:, then the SVM based speaker identification will recognize Speaker A by the input speech utterance, and the speech identification system will recognize the Speaker B by the same word string of speech utterance.

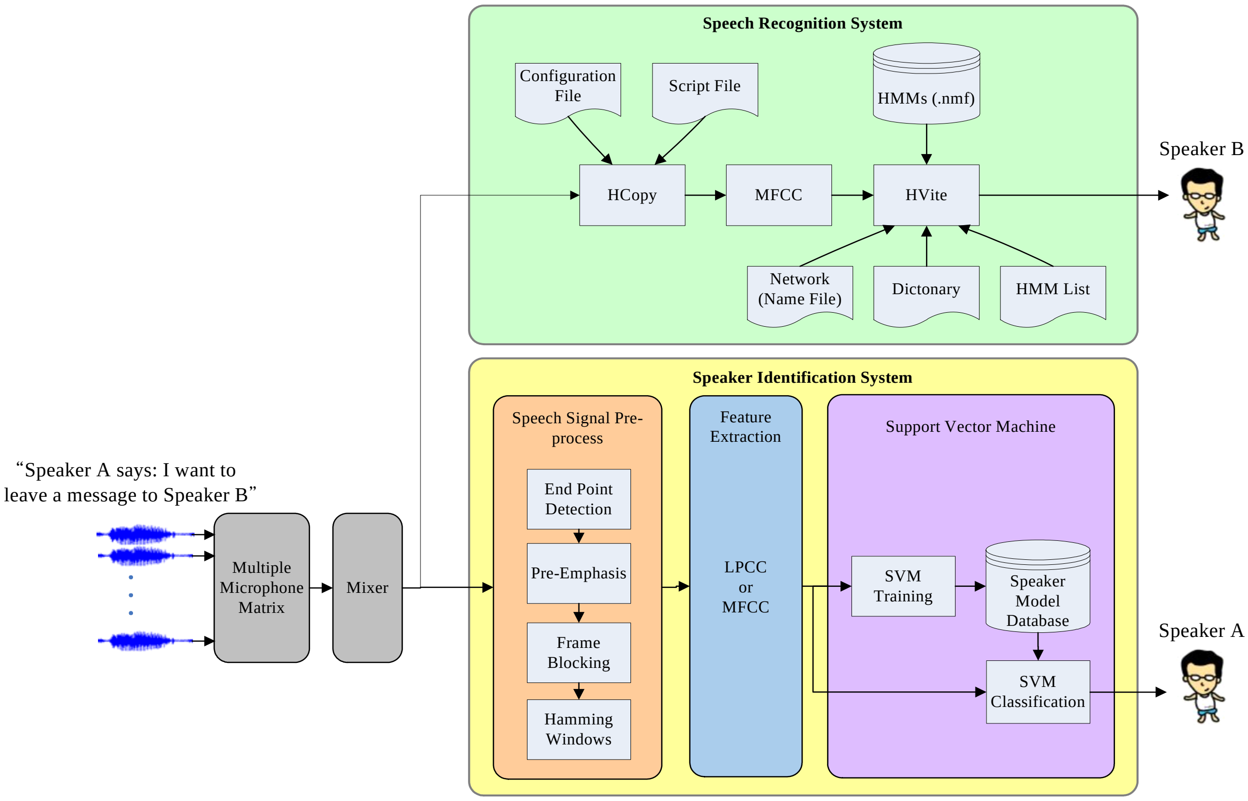

Fig.3.1 Proposed Consolidation of Speaker and Speech Recognition System

3.2 Feature Extraction and VAD Process of Consolidated System

The input speech signal of the consolidated speech system is collected from the ubiquitous speech environment [6]. The ubiquitous speech environment is a microphone-array framework which is composted of six microphones on the far-field space, when the speech signal are received then mixed to be only one speech signal by mixer and output for feature extraction. The 18th order of LPCC feature extraction method is for SVM based speaker identification, while MFCC features are obtained by using HTK for speech recognition system.

Based on utilizing the information of energy, zero-crossing rate and the spectral flatness, then the informational signal frame is thus detected and non-information frame is ignored. If the silence is presented among two or more frames between signals, then that particular segments must be removed from the original speech signal.

VAD (Voice activity detection) method of EPD (end-point detection) is also adopted. The goal of EPD is to identify the important part of informational signals for further processing. The EPD is also known as "speech detection" or "voice activity detection", which usually play an important role in speech signal processing and recognition.

3.3 Multi-class SVMs of Testing Phase in Speaker Identification

The Multi-class (C classes) training process is to train C(C-1)/2 different 2-class SVMs on all possible pairs of classes, and then to classify test points according to which has the highest number of ‘votes’, this approach is called one-versus-one [12]. Table2.3 are shown that all possibly pair-wise classifier and listed as diagonal symmetric matrix, each pair-wise classifier is defined as a hyper plane between two classes. The repetitions of the pair-wise are presented from the diagonal symmetric matrix. According to the analysis and utilize the diagonal pair-wise, then only C(C-1)/2 “votes” , which also means that there are C(C-1)/2 hyperplanes are acquired for comparison.

Table 2.3. The method of C-classes for SVM training process

~~1 vs. 1~~	1 vs. 2	1 vs. 3	…	1 vs. C
~~2 vs. 1~~	~~2 vs. 2~~	2 vs. 3	…	2 vs. C
~~3 vs. 1~~	~~3 vs. 2~~	~~3 vs. 3~~	…	3 vs. C
~~C vs. 1~~	~~C vs. 2~~	~~C vs. 3~~	~~…~~	~~C vs. C~~

In testing phase, the extracted input sample using discriminate function to make a decision that whether the input sample are resided in Class 1 or Class 2. If the result of discriminate function y≧+1, then input data belongs Class 1, otherwise if y≦-1 then it is in Class 2. Eq.(7) is shown that symbol of $vote_{1,2}$ is the normalized result of input samples in Class 1, while $vote_{1,2}$ is the testing sample in Class 2.

$$
\begin{aligned}
vote_{1,2} &= \frac{test\ \ pattern\ in\ calss\ 1}{N} \\
vote_{2,1} &= \frac{test\ \ pattern\ in\ calss\ 2}{N} \\
vote_{1,2} + vote_{2,1} &= 1
\end{aligned}
\qquad (7)
$$

Based on Eq.(7), when the input sample data after undergoing C(C-1) hyper plane or *vote*$_{i,j}$ computations except diagonal *vote*$_{i,i}$ in Table.2.4. In order to find the high score of the target speaker, then to summarize *votes* from each row beside diagonal votes in Table.2.4. And then to compare values of summarized rows and to find out the maximum through each class. Finally, the maximum value of the class is represented the target speaker. The corresponding equation of evaluating target speaker is shown in Eq.(8).

Table 2.4. The *vote*$_{i,j}$ value of each class

Compare Class j Value of type i	Class 1	Class 2	Class 3		Class C
Class 1	$vote_{1,1}$	$vote_{1,2}$	$vote_{1,3}$	…	$vote_{1,C}$
Class 2	$vote_{2,1}$	$vote_{2,2}$	$vote_{2,3}$	…	$vote_{2,C}$
Class 3	$vote_{3,1}$	$vote_{3,2}$	$vote_{3,3}$	…	$vote_{3,C}$
:	:	:	:	:	:
Class C	$vote_{C,1}$	$vote_{C,2}$	$vote_{C,3}$	…	$vote_{C,C}$

$$
\begin{aligned}
&class\ i\ value\ is\ vote_i = \sum_{j=1}^{C} vote_{i,j} \\
&target\ class = \arg\max_i vote_i
\end{aligned}
\qquad (8)
$$

3.4 HTK based Speech Recognition for Testing Phase

The components of proposed consolidated system in speech recognition are included as 1) Tree lexicon, 2) The task grammar, and 3) Viterbi beam search [10]. The first component is to create a dictionary. The dictionary provides an association between words used in the task grammar and the acoustic models, in that may be composed of sub word (phonetic, syllabic etc.) units. The second component of task grammar is to constrain on what the recognizer can expect as input, as the system built, then a voice operated interface is provided for name recognition, it is capable of handling the word strings. In order to limit the scope of this work, only the syllable to deal with name grammars is needed. The final component is the Viterbi beam search. This component is essentially a dynamic programming algorithm, consisting of traversing a network of HMM states and maintaining the best possible path score at each state in each frame. It is a time-synchronous search algorithm in that it is to process all states completely at time *t* before moving on to time *t + 1*.

After three components are finished, and the recognizer is complete and ready for evaluating the performance. The recognition process can be summarized as in the top part of Fig.3.1 related to Speak B. In the beginning, the input speech signal is transformed into a series of”acoustical vectors” (here MFCCs) by using the HTK tool HCopy, in the same way as what was done with the training data. The input observation is then processed by the Viterbi algorithm using the HTK tool HVite.

4. Experimental Results

In order to evaluate the performance of the consolidated system in real life, thus the experiments are tested in the Aspire Home, which is located in the NCKU Chi-Mei Building. The training and testing phase of the individual speaker identification system and speech recognition system as well as proposed consolidated speech recognition system are setup and evaluated, respectively.

4.1 Experimental Results of Individual Speaker Identification System

The component of speaker identification system is to use 18 order of LPCC feature. Ten seconds of speech utterances is for training, and two second of speech utterance is for testing. Total 10 persons are assessed in this case. The key elements of speech preprocess include the end point detection and voice activity detection. The parameters γ and C are setting to be 0.0005 and 50, respectively. The sample rate is 16 kHz, and the frame size is 512 points. Three types of microphone configuration is to be assessed, such as single microphone, wireless microphone and microphone array. Two modes of background noise are also adopted, i.e. silence mode

	Single Microphone (Omni-directional)	Wireless Microphone (Omni-directional)	Microphone Array (Omni-directional)
Accuracy Rate (Silence Mode)	76.6%	91.6%	96.6%
Accuracy Rate (TV noise Mode)	66.7%	86.6%	73.3%

Fig. 4.1. The experimental results of individual component in speaker Identification system

v.s. TV noise mode are built in test environment. The experimental results of individual component in speaker identification system are shown in Fig. 4.1

4.2 Experimental Results of Individual Speech Recognition System

The training databases include English Across Taiwan (EAT) and Mandarin Across Taiwan-400 (MAT-400). Using man-made sifting way, then EAT are totally remaining 8375 wave files, including English long sentences, English short sentences and English words. The corpus contains 19221 words for training. In MAT-400, the MATDB-4 (1200) and MATDB-5 (400) category are adopted. By using man-made sifting way, there are totally remaining 15400 wave files, including words of 2 to 4 syllables and phonetically balanced sentences. The corpus contains 80903 words for training. There are one hundred of testing word strings, which can be regarded as voice command, and the content included Chinese movie name, English words and Chinese/English mixture sentence and etc. There are totally 10 people to test this system. The speaker randomly selected twenty sentences of testing string to evaluate the system. The experimental results of individual component in speech recognition system are shown in Fig. 4.2

Number of speaker	Positive Identification	Negative Identification	Testing Times	Accuracy Rate
8 male speakers	127	33	160	79.38%
2 Female speakers	24	16	40	60.00%
Total Speaker	151	49	200	75.5%

Fig.4.2. the experimental results in speech recognition system

4.3 Experimental Results of Consolidated System

The consolidated speech recognition system is examined in the real doorway of Aspire House in NCKU. In this experiment, totally six persons are joined test. The training and testing utterance period is the same as speaker identification system. In the speech recognition, there are six sentences of word string for testing, each testing pattern is formatted as “I want to leave message to XXX”, the sub-string “XXX” is represented as the name of six speakers in English or Mandarin. When the speaker pronounces the randomly selected testing pattern within six sentences, the system will identify two target speakers at the same time from the real speaker and word string speaker. Each speaker is to test the system for thirty times. In this test, the single microphone configuration is used for assessment. The experimental results of consolidated system are shown in Fig. 4.3.

	Positive Identification	Negative Identification	Consolidated System Accuracy
Speaker Identification Accuracy	157	23	87.22%
Speech Recognition Accuracy	164	16	91.1%

Fig.4.3. the experimental results of consolidated speaker and speech recognition system

5. Conclusion

For human-centric digital life, this paper has presented an integrated architecture of speaker identification and speech recognition for intelligent doorway application. This framework is capable of recognize two target speakers via only one-word utterance. The SVM is used for speaker identification and the confusion matrix is used for develop the multi-lingual speech recognition system. This integrated system has been realized in the intelligent doorway of our prototype digital house. The future work intends to integrate a sound localization technique to localize the speaker position.

References

[1] V. N. Vapnik, Statistical Learning Theory. New York: Wiley, 1998.

[2] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition,” Data Mining and Knowl. Discov., vol. 2, no. 2, pp. 1–47, 1998.

[3] V. Wan and S.Renals, ”Speaker verification using sequence discriminant support vector machines,” IEEE trans. On Speech and Audio Processing vol.13, no. 2, Mar.2005.

[4] Q. Jin, T. Schultz, and A. Waibel, “Far-Field Speaker Recognition,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 7, Sep. 2007.

[5] M. Ji, S. Kim, H. Kim, K. C. Kwak, and Y.J. Cho, “Reliable Speaker Identification Using Multiple Microphones in Ubiquitous Robot Companion Environment,” 16th IEEE International Conference on Robot & Human Interactive Communication.

[6] J. F. Wang, T.W. Kuan, J.C. Wang, and G. H. Gu,”Ubiquitous and Robust Text-Independent Speaker Recognition for Home Automation Digital Life,” UIC 2008, LNCS 5061, pp. 297–310, 2008.

[7] C.L. Huang, C-H Wu, “Generation of Phonetic Units for Mixed-Language Speech Recognition Based on Acoustic and Contextual Analysis”, Department of

Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, R.O.C. (2007)

[8] C. Y MA, Pascale FUNG, "Using English Phoneme Models for Chinese Speech Recognition" , The Human Language Technology Center Department of Electrical and Electronic Engineering Hong Kong University of Science and Technology (HKUST), Hong Kong

[9] F. Seide, J. C. Wang, 1998. Phonetic modeling in the Philips Chinese continuous-speech recognition system. In Proc.

[10] P.Y. Shih, J.F. Wang, H.P. Lee, H.J. Kai, H.T. Kao, Y.N. Lin ," Acoustic and Phoneme Modeling Based on Confusion Matrix for Ubiquitous Mixed-Language Speech Recognition," IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing, DOI 10.1109/SUTC Jun,2008

[11] J.C. Wang, C.H.Yang, J.F. Wang, and H.P. Lee, "Robust speaker identification and verification," IEEE Compu. Intell. Mag., pp.52-59, May 2007.

[12] C. M. Bishop, Pattern Recognition and Machine Learning, New York, NY :Springer Science+Business Media, 2006, pp. 325-358

[13] J.C. Platt," Sequential Minimal Optimization for SVM," Published by Pennsylvania State University, http://citeseerx.ist.psu.edu/. 2007

[14] J. F. Wang, and G. H. Gu," Ubiquitous and Robust Text-Independent Speaker Recognition and FPGA Implementation for SMO algorithm of SVM", Master Degree Dissertation, July, 2008.