PHONE SET GENERATION BASED ON ACOUSTIC AND CONTEXTUAL ANALYSIS FOR MULTILINGUAL SPEECH RECOGNITION

Chien-Lin Huang and Chung-Hsien Wu

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, R.O.C.

ABSTRACT

This study presents a novel approach to generating phone units generation for the recognition of multilingual speech. Acoustic and contextual analysis is performed to characterize multilingual phonetic units for phone set generation. A confusion matrix combining acoustic and contextual similarities between every two phonetic units is constructed for phonetic unit clustering. Acoustic likelihood and hyperspace analog to language (HAL) model are adopted for acoustic similarity and contextual similarity estimation of phone models, respectively. Experiments show that the generated phone set provides a compact and robust set that considers acoustic and contextual information for multilingual speech recognition.

Index Terms—Multilingual speech recognition, confusion matrix, acoustic likelihood, hyperspace analog to language model

1. INTRODUCTION

Speech is the most convenient medium for human-to-human communication and human-to-machine interaction. Automatic speech recognition is vital for many applications, including speech summarization, spoken document retrieval and speech dictation. Due to the trend of globalization, multilingualism frequently occurs in speech content, and the ability of speech recognition systems to process speech in multiple languages has become increasingly desirable.

Multilingual speech recognition can be achieved in various ways. One approach employs external language identification (LID) systems [1] to identify the language of the input utterance. The corresponding monolingual system is then selected to perform the speech recognition [2]. A multilingual system produced using this approach achieves the same performance as the monolingual systems when no LID errors occur. Accuracy of the external LID system is the main concern in overall system performance. Recently, approaches to multilingual speech recognition focus on the utilization of a multilingual phone set. The first and simplest approach to phone set definition is to combine the phone inventories of different languages together without sharing the units across the languages, but doest not share parameters across the languages in the models. The size of the multilingual phone inventory will increase proportionally to the number of languages in the multilingual recognition system.

The second one is to map the language-dependent phone set to a global inventory of the multilingual phonetic associations based on phonetic knowledge to construct the multilingual phone inventory. Several global phone-based phonetic representations such as the International Phonetic Alphabet (IPA) [3], Speech Assessment Methods Phonetic Alphabet [4] and Worldbet [5] are generally used. The advantage is that the multilingual phone symbols have clear representation in the context. However, only the phonetic knowledge, rather than statistical similarity measure is employed. The direct IPA mapping does not consider the spectral properties of the phone models. Hence, acoustic model parameters cannot precisely describe the distribution of the real training data.

The third one is to merge the language-dependent phones using a hierarchical phone clustering algorithm to obtain a compact multilingual phone inventory. In this approach, the distance measure between acoustic models, such as Bhattacharyya distance [6], Mahalanobis distance or Kullback-Leibler divergence [7], is employed to perform bottom-up clustering. The advantage of this approach is that the distance is estimated from the statistical similarity measure of real recognition models. However, phone clustering is performed based on acoustic features rather than the likelihood of phone models.

This study presents an approach to phonetic unit generation for multilingual speech recognition. Context-dependent triphones for Mandarin and English speech are constructed based on the IPA representation. Acoustic and contextual analysis is performed to characterize the multilingual context-dependent phonetic units. A data-driven approach based on a multilingual corpus statistically determines the confusing characteristic among phonetic units given the matrices from acoustic and contextual similarities. Finally, the modified *k*-means algorithm is adopted to cluster the context-dependent phonetic units to obtain a compact and robust phone set.

2. MULTILINGUAL PHONE DEFINITION BASED ON ACOUSTIC AND CONTEXTUAL ANALYSIS

This study uses IPA-based multilingual phone definition, which is exploited for phonetic representation. Using phonetic representation of the IPA can lower the number of recognition units in multilingual speech recognition applications. Considering the co-articulated pronunciation, context-dependent triphones, expanded from the IPA-based phonetic units, are adopted.

In multilingual speech recognition, misrecognition of phones is caused by the incorrect pronunciation or the confusable phonetic set. This study focuses on the confusing characteristic of multilingual phonetic set. The statistical methods are proposed to deal with the problem of misrecognition caused by the confusion between phonetic units in multilingual speech recognition. Based on the analysis of confusing characteristics, several confusing phones caused by the confusable phonetic representation can be redefined.

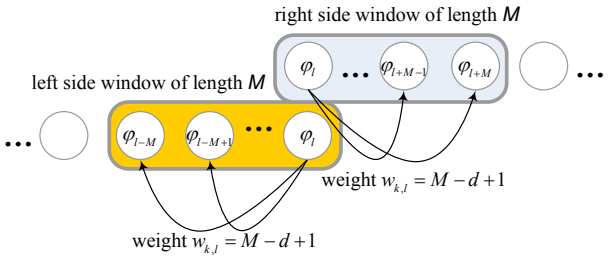

Figure 1. The weighting scheme of the HAL model

2.1. Confusion Matrix Based on Acoustic Likelihood

The posterior probabilities obtained from the phone-based hidden Markov model (HMM) are employed. Given two phone models, φ_k and φ_l, trained with the corresponding training data, $x_i^k, 1 \le i \le I$ and $x_j^l, 1 \le j \le J$, the symmetric acoustic likelihood (ACL) between phone models, φ_k and φ_l, is estimated as follows

$$a_{k,l} = \frac{\sum_{i=1}^{I} P(x_i^l \mid \varphi_k) + \sum_{j=1}^{J} P(x_j^k \mid \varphi_l)}{I + J} \tag{1}$$

where I and J represent the number of training data for phone models, φ_k and φ_l, respectively. The acoustic confusion matrix $\mathbf{A} = (a_{k,l})_{N \times N}$ is obtained from the pairwise similarities between every pair of phone models, and N is the number of triphones.

2.2. Confusion Matrix in Contextual Analysis

A co-articulation pattern can be considered as a semantically plausible combination of phones. This study adopts text mining method to induce co-articulation patterns automatically from a multilingual speech corpus. A crucial step to inducing the co-articulation patterns is to represent speech intonation as well as combinations of phones. To reach this goal, the hyperspace analog to language (HAL) model [8] is adopted to construct a high-dimensional contextual space for the multilingual speech corpus. Each context-dependent triphone in the HAL space is represented as a vector of its contextual information, which indicates that the sense of a phone can be co-articulated through its neighboring phones. This notion is derived from the observation of articulation behavior. Based on the co-articulation behavior, if two phones share more common neighboring phones, then they are more similarly articulated.

In HAL space, each vector dimension is a weight representing the strength of the association between the target phone and its neighboring phone. The weights are computed by applying an observation window of length M over the corpus. Every phone within the window is considered as part of the pronunciation co-articulated with every other phone. The weight between two phones of distance d within the window is given by $M - d + 1$. Figure 1 shows the weighting scheme of the HAL model [8]. The HAL space $\mathbf{G} = (g_{k,l})_{N \times N}$ is constructed after moving the window by one phone increment over the sentence. The resultant HAL space is an $N \times N$ matrix.

Table 1 presents the HAL space for the English and Mandarin mixed sentence "查一下<look up> (CH A @ I X I A) Baghdad (B AE G D AE D)."

Table 1. Example of multilingual sentence in HAL space

	CH	A	@	I	X	B	AE	G	D
CH									
A	3			4	1				
@	2	3							
I	1	2	4		3				
X		1	2	3					
B		3		2	1				
AE		2		1		3		2	3
G		1				2	3		
D						1	5	4	

The row vector of each phone in Table 1 represents its left neighboring phones, i.e. the weights of the phones preceding it. The corresponding column vector represents its right neighboring phones. $w_{k,l}$ denotes the k^{th} weight of the l^{th} triphone φ_l. Furthermore, the weights in the vector are re-estimated as follows

$$\overline{w}_{k,l} = w_{k,l} \times \log \frac{N}{N_l} \tag{2}$$

where N_l denotes the number of vectors of phone l with nonzero dimension. The HAL space is transformed into a probabilistic framework after each dimension is re-weighted. Each weight can thus be redefined as

$$\hat{w}_{k,l} = \frac{\overline{w}_{k,l}}{\sum_{k}^{N} \overline{w}_{k,l}} \tag{3}$$

In order to generate a symmetric matrix, the weight is averaged as

$$g_{k,l} = \frac{\hat{w}_{k,l} + \hat{w}_{l,k}}{2} \quad 1 \le k, l \le N \tag{4}$$

2.3. Phone Clustering

For phone clustering, the confusion matrix $\mathbf{V} = (v_{k,l})_{N \times N}$ contains pairwise similarities between every two multilingual triphones computed as

$$v_{k,l} = -(\beta \times \log(a_{k,l}) + (1 - \beta) \times \log(g_{k,l})) \quad 1 \le k, l \le N \tag{5}$$

where k and l represent the elements of row and column in the matrix. $\beta = 0.1$ denotes the combination weight. This study calculates the confusing characteristics from probability to distance with the log operator. The sum rule of data fusion [9] is applied to combine the confusion matrices of ACL and HAL. This study expects to cluster the context-dependent triphones with

similar acoustic and contextual properties into a multilingual triphone cluster. The modified *k*-means (MKM) algorithm [10] is applied to cluster the confusing triphones into a phonetic unit using the cosine measure between triphones. The convergence of closeness measure is determined by a preset threshold.

3. EXPERIMENTS

3.1. Description of the HMM-Based ASR System

An in-house multilingual speech recognizer was implemented for experiments. Continuous HMMs were adopted for acoustic modeling. The Gaussian mixture number per state of the acoustic HMMs ranges from two to 16, depending on the quantity of the training data. The silence model was a one-state HMM with 32 Gaussian mixtures trained with the non-speech segments. The multilingual speech recognition employed the tree copy search decoding algorithm [11]. Figure 2 displays an example of the lexical pronunciation tree, using the recognized multilingual sentence "Disney (D, IH, Z, N, I) 樂園<Land> (L, E, V, AN)" depicted in a simplified schematic form. The vocabulary size of English and Mandarin for the speech recognizer was 4,271 words in the experiments.

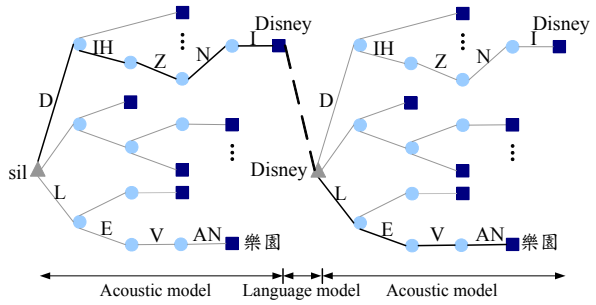

Figure 2. The multilingual tree copy search decoding. This lexicon tree starts with silence (sil). The "▲" is the root of a tree copy for the previous history. Line " — " is the acoustic model recombination within a tree copy. Line "– –" is language model recombination. The "■" is the word end hypothesis.

3.2. Multilingual Database

In Taiwan, English and Mandarin are popular in daily conversation, culture, and media. The English across Taiwan project (EAT [online] http://www.aclclp.org.tw/) sponsored by National Science Council, Taiwan, designed 600 recording sheets for bilingual corpus collection. Each sheet contains 80 reading sentences, including English long sentences, English short sentences, English words and mixed English and Mandarin sentences. Each sheet was pronounced by the English Department students and non-English Department students for speech recording. The corpus was recorded as sound files with 16 kHz sampling rate and 16 bit resolution. Table 2 shows the information of the collected corpus.

Table 2. EAT multilingual corpus information

	English Department		non-English Department	
	male	female	male	female
Sentence	11,977	30,094	25,432	15,540
Person	166	406	368	224

3.3. Evaluation of the Multilingual Phone Generation

In order to evaluate the performance of the generated multilingual phone set, three classes of phone-based recognition errors, namely insertion errors (*ins*), deletion errors (*del*) and substitution errors (*sub*), were considered. The phone recognition accuracies for (*acc*) was estimated as

$$acc = \frac{num - ins - del - sub}{num} \times 100\% \quad (7)$$

where *num* denotes the number of phones in the test database. The phone accuracy of different approaches as a function of the number of states in HMMs is shown in Fig. 3.

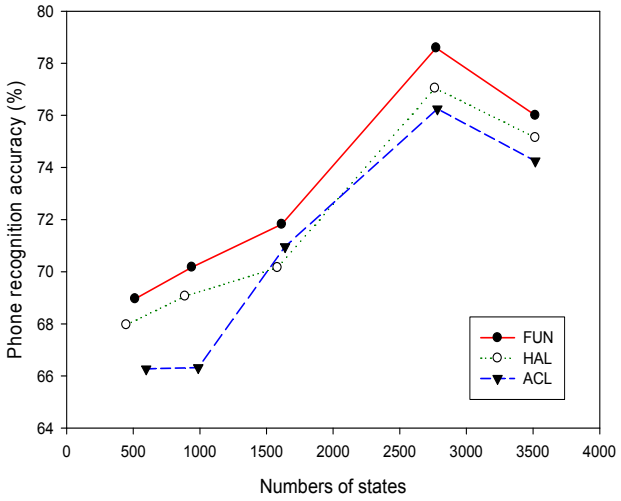

Figure 3. Phone recognition accuracies of ACL, HAL and FUN methods for the total number of distinct HMM states.

Three different approaches including acoustic likelihood (ACL), contextual analysis (HAL) and fusion of acoustic and contextual analysis (FUN) were adopted for experiments. The proposed fusion method clearly achieved a better result than the individual ACL and HAL methods. HAL achieves a higher recognition rate than ACL in acoustic analysis and contextual analysis revealing that the contextual analysis is more significant than acoustic analysis for multilingual phone clustering. The curves in the plot show that the phone accuracy increases with the increasing in number of states, and then finally decreases due to the confusing triphone definition and the requirement for a large size of multilingual training corpus. The proposed multilingual phone generation approach can obtain an improved performance over the ordinary multilingual triphone set. In this study, the multilingual speech recognition system for English and Mandarin contains 924 context-dependent triphone models.

3.4. Comparison of Acoustic and Language Models for Multilingual Speech Recognition

The phonetic units of Mandarin and English can be represented by 37 and 39 fundamental phones, respectively. For the comparison of monophone and triphone-based recognition, different phone inventories shown in Table 3 are considered. The monophone approach adopts direct combination of Mandarin and English language-dependent phones (MIX) and language-dependent IPA phones (IPA). The triphone methods (TRE) are generated with the use of a phonetic tree-based clustering procedure, suggested by [12] and our proposed method (FUN).

Table 3. Comparison of acoustic and language models for multilingual speech recognition

	Monophone		**Triphone**	
	MIX	IPA	TRE	FUN
With language model	45.81%	66.05%	76.46%	78.18%
Without language model	32.58%	51.98%	65.32%	67.01%

In acoustic comparison, multilingual context-independent (MIX and IPA) and context-dependent (TRE and FUN) phone sets were investigated. With the language model of English and Mandarin, the MIX approach achieved 45.81% phone accuracy and the IPA method achieved 66.05% phone accuracy. The performance of the IPA method is evidently better than that of the MIX approach. TRE method achieved 76.46% phone accuracy and our proposed approach achieved 78.18%. It is obvious that the approach using triphone models achieves better performance than that using monophone models. There is around 2.25% relative improvement from 76.46% accuracy for the baseline system based on TRE to 78.18% accuracy for the approach using acoustic and contextual analysis.

In order to evaluate the performance of acoustic modeling, experiments were conducted without considering the language model. Without the language model of English and Mandarin, the MIX approach achieved 32.58%, IPA method achieved 51.98%, TRE method achieved 65.32% and the proposed approach achieved 67.01% phone accuracies.

3.5. Comparison of Monolingual and Multilingual Speech Recognition

In this experiment, utterances of English words and sentences in the EAT corpus were collected for the evaluation of monolingual speech recognition. Table 4 presents a comparison of monolingual and multilingual speech recognition using the EAT corpus

Table 4. Comparison of monolingual and multilingual speech recognition

	Monolingual		**Multilingual**
	English word	English sentence	English and Mandarin mixed sentence
Phone recognition accuracy	76.25%	67.42%	67.01%

For context-dependent phone modeling without language model, the accuracy of monolingual English word achieved 76.25% which is higher than the 67.42% for monolingual English sentences. The phone recognition accuracy of monolingual English sentences was 67.42%. The performance is slightly better than the 67.01% for mixed English and Mandarin sentences

4. CONCLUSION

This paper proposes acoustic and contextual analysis to generate phonetic units for multilingual speech recognition. Context-dependent triphones are defined based on the IPA representation and the confusing characteristics of the multilingual phone set are analyzed using acoustic and contextual information. In acoustic analysis, the acoustic likelihood confusion matrix is constructed by the posterior probability of triphones, and in contextual analysis, the HAL approach is employed to model the contextual property. Based on the confusion matrix representing the pairwise phone similarities, the modified *k*-means algorithm is used to cluster the multilingual triphones into a compact and robust phone set. Experimental results show that the generated phone set gives an encouraging improvement on recognition performance.

5. REFERENCES

[1] Chung-Hsien Wu, Yu-Hsien Chiu, Chi-Jiun Shia, and Chun-Yu Lin, "Automatic Segmentation and Identification of Mixed-language Speech using Delta-BIC and LSA-based GMMs," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 266-276, 2006.

[2] Alex Waibel, Hagen Soltau, Tanja Schultz, Thomas Schaaf, and Florian Metze, "Multilingual Speech Recognition. Chapter in Verbmobil: Foundations of Speech-to-Speech Translation," Springer-Verlag, 2000.

[3] Mathews, R. H., Mathews' Chinese-English Dictionary, Caves, 13th printing, 1975.

[4] J. C. Wells, Computer-Coded Phonemic Notation of Individual Languages of the European Community. J. IPA, 19, pp. 32-54, 1989.

[5] James L. Hieronymus, ASCII Phonetic Symbols for the World's Languages: Worldbet. Journal of the International Phonetic Association, 1993.

[6] Brian Mak and Etienne Barnard, "Phone clustering using the Bhattacharyya distance," *in Proc. ICSLP*, pp. 2005-2008, 1996.

[7] Jacob Goldberger and Hagai Aronowitz, "A Distance Measure Between GMMs Based on the Unsented Transform and its Application to Speaker Recognition," *in Proc. of EUROSPEECH*, pp. 1985-1988, Lisbon, Portugal, 2005.

[8] D. Song and P.D. Bruza, "Towards context sensitive information inference," *Journal of the American Society for Information Science and Technology*, vol. 54, no. 4, pp. 321-334, 2003.

[9] Josef Kittler, Mohamad Hatef, Robert P.W. Duin, and Jiri MatasOn, "On Combining Classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226-239, 1998.

[10] Jay G. Wilpon and Lawrence R. Rabiner, "A modified K-means clustering algorithm for use in isolated work recognition," *IEEE Transactions on Acoustics, Speech, and Signal Proc.*, vol. 33, no. 3, pp. 587-594, 1985.

[11] Hermann Ney and Stefan Ortmanns, "Progress in dynamic programming search for LVCSR," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1224-1240, 2000.

[12] S.J. Young, J.J. Odell, P.C. Woodland, "Tree-based State Tying for High Accuracy Acoustic Modelling," *in Proc. ARPA Human Language Technology Conference*, Plainsboro, USA, 1994.