An Experimental Study on Continuous Phone Recognition with Little or No Language-Specific Training Data

Dau-Cheng Lyu[1], *Sabato Marco Siniscalchi*[2] *and Chin-Hui Lee* [3]

[1] Department of Electrical Engineering, Chang Gung University, Tao-Yuan, Taiwan
[2] Department of Electronics and Telecommunications
Norwegian University of Science and Technology, Trondheim, Norway
[3] School of ECE, Georgia Institute of Technology, Atlanta, GA 30332 USA
d9221003@stmail.cgu.edu.tw, marco77@iet.ntnu.no, chl@ece.gatech.edu

Abstract

We study continuous phone recognition with little or no language-specific speech training data. The phone recognizer integrates three levels of information from: (1) frame based speech attribute detectors, (2) artificial neural network based phone event mergers, and (3) decoding based evidence verifiers. With a set of acoustic phonetic attributes defined over a number of available languages, a collection of attribute-to-phone mapping rules can either be specified in a language-dependent way, one for each language, or even independently for all languages if the attribute specification is complete to cover all phones and the phone definition is universal to cover all spoken languages. We report on experimental results on Japanese phone recognition with the OGI Multilingual Speech Corpus. It is interesting that a good performance can be achieved without using any Japanese speech training data, and the phone accuracy rates vary depending on how the attribute detectors and phone mergers are configured. Further improvement is observed by adding little Japanese data to train the attribute-to-phone mergers.

Index Terms: multilingual speech recognition, detection-based speech recognition, speech attribute detection, attribute to phone mapping, phone model with little or no training data

1. Introduction

Automatic speech recognition (ASR) is a process to convert speech into a sequence of words. Top ASR performance is often delivered by using a large amount of language-specific speech data to train the set of language-specific acoustic models, one for each language. Although such speech examples are available for resource-abundant languages, such as English, Arabic and Mandarin, this scenario does not always apply to all the languages under consideration. Developing ASR systems for resource-insufficient languages has been a research topic of recent interest [1].

Multi-lingual speech recognition is a popular approach to dealing with the above problem [1, 2]. In such a system fundamental speech units with similar sounds across different languages are grouped together and represented by a single phonetic symbol. The set of such collection of symbols are often referred to as a universal phone set. The International Phonetic Alphabet (IPA) [3] is such an attempt to define a set of universal phoneme units that can be used to represent all speech sounds for all spoken languages.

By collecting a large set of speech examples covering all speech units and their contexts, a set of acoustic models can be trained to represent all units needed in any language so that an ASR system can be designed even for languages with no speech samples available in the general training corpus.

Another way to model a universal phone set to cover all spoken languages is through defining a set of acoustic phonetic attributes so that all the phones can be modeled by merging information from a small number of speech attributes. Since not all phones in a universal phone set are present in a particular language, it requires more languages to collectively define a phone set to cover all languages. On the other hand, most fundamental speech attributes, such as voicing, nasality, dental, and frication, can be identified from a particular language. Therefore it is believed that such attributes can be shared across many different languages. A set of 21 acoustic phonetic attributes was used in a detection-based ASR study [4]. Continuous phone recognition is performed in a bottom-up manner by integrating three levels of information from: (1) frame based speech attribute detectors, (2) artificial neural network based phone event mergers, and (3) decoding based evidence verifiers.

The above three modules were part of a detection-based framework, called automatic speech attribute transcription (ASAT), recently proposed for developing ASR techniques and systems [5]. For Modules 1 and 2, the attribute detectors and attribute-to-phone mapping mergers can be trained with data-driven techniques, while for phone recognition Module 3 can be performed by a conventional Viterbi decoder.

It was shown in [5] that the same set of 21 attributes can be used across six different languages, namely English, German, Hindi, Japanese, Mandarin, and Spanish, for continuous phone recognition. The 21 detectors in Module 1 serve as a way to compute posterior attribute probabilities and can be trained with language-specific data, or with multilingual speech data from all six languages. While for Module 2, language-specific mapping rules have been implemented with artificial neural networks (ANNs) to combine information from a small set of attributes used to define the phone. Cross-language attribute-to-phone mapping ANNs have also been experimented briefly in [5]. They are more difficult to train for at least the following two reasons: (1) some phones appearing in a particular language are often not seen in other languages, and therefore such mapping ANNs can not be trained without seeing some corresponding phone-specific speech samples; and (2) the phone definition used in a universal phone set is often not consistent across multiple languages, i.e. one common symbol used in two or more languages may require separate symbols, or different phone symbols in multiple languages may share the same fundamental speech unit of interest.

This work expands upon our pervious study on cross-language experiments [5] and presents additional results on

different ways of using the limited amount of information available to us for training the attribute detectors and phone mergers with little or no language-specific training data. In particular, we focus on the following three research issues: (1) the effect on the phone recognition performance of a target language when changing the amount of training materials for designing detectors and mergers; (2) the effect of designing cross-language attribute-to-phone mapping ANNs using various language training materials with different phone sets; and (3) designing of language-specific phone mapping rules by grouping rules available from other languages. All these issues can only be investigated because a common set of fundamental speech attributes is defined and share across all the languages under consideration.

In this study, the target language is chosen to be Japanese which is the probably the easiest in the six-language OGI Multilingual Telephone Speech Corpus because the size of the Japanese (JAP) phone set is relatively small and the phone definition is rather clear for defining and designing cross-language attribute-to-phone mapping rules. We use the other five languages in the OGI Corpus, namely English (ENG), German (GER), Hindi (HIN), Mandarin (MAN) and Spanish (SPA), as the available training data for training detectors and mergers for Japanese without using any Japanese speech data. We also experiment with adding some Japanese speech utterances to improve the attribute-to-phone mergers. It is interesting to note that we approach language-specific phone recognition accuracy of a target language without using any training speech data from the target language. By adding a small amount of language-specific speech training data, the phone recognition performance can further be improved. We believe this detection-based multilingual speech recognition approach is a promising research direction to pursue for designing ASR systems for less-resourceful languages.

2. System Overview

As proposed earlier [4, 5] we use the same detection-based phone recognition system shown in the block diagram in Figure 1 consisting of three main modules: (1) a bank of speech attribute detectors, (2) a set of phone mergers, and (3) an evidence verifier. More detail about each module relevant to the current study will be provided in the following sections. The evidence verifier generates only the first best hypothesis as the recognized phone string.

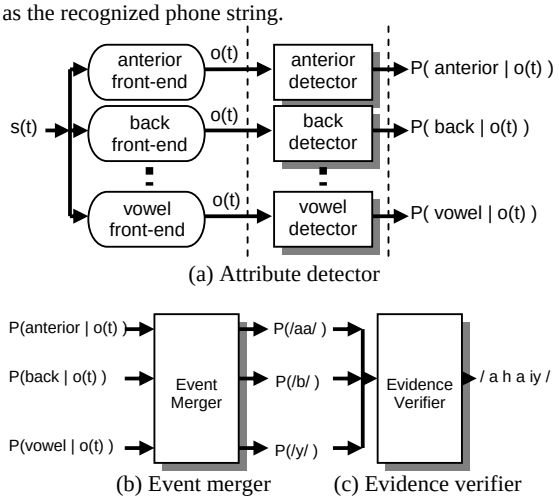

Figure. 1. Overall system

2.1. Speech Attribute Detectors

Table 1 lists the set of phonological features that we use in our experiments. We refer to these phonetic features as attributes.

Manner	approximant, fricative, nasal, vowel, stop
Place	anterior, back, continuant, coronal, dental, glottal, high, labial, low, mid, retroflex, round, silence, tense, velar, voiced

Table 1. Speech attributes.

The main purpose of each attribute detector is to analyze speech and produce a confidence or posterior probability score that pertains to the acoustic phonetic attribute of interest. We build each detector using 3 feed-forward ANNs with one hidden layer of 500 hidden nodes as organized in [6]. To estimate the ANN parameters, we separate the training data into attribute present and attribute absent regions for every event of interest using the available phonetic transcription from a training corpus. A softmax activation function is used at the output layer to produce an approximate posterior probability that a speech particular event appears at the frame currently being processed. Energy trajectories in mel-frequency bands, organized in a split-temporal context as in [7], are used as the parametric representations of speech.

2.2. Event Mergers and Evidence Verifiers

An attribute-to-phone merger produces a frame-level phone score by combining together detector outputs from attributes corresponding to the phone of interest with different weights. All phone mergers are implemented using a single feed-forward ANN with one hidden layer of 800 hidden nodes. The softmax function is again used at the output layer.

The evidence verifier is just a decoding network which consists of a set of context independent phone models layered in parallel and with uniform entrance probabilities. Each phone is modeled by a 3-state left-to-right hidden Markov model (HMM) [8]. The HMM state likelihood is the phone posterior probability provided by the corresponding phone merger. We assume equal prior probabilities for all phones. A Viterbi algorithm is performed over the decoding network to generate the decoded sequence of phones.

3. Experiments and Results

We adopted the "stories" part of the OGI Multi-language telephone speech corpus [9] for all experiments. Phonetic transcription is available for all six languages. For each language, we divided the available speech data into three subsets, namely: training, validation, and test. Table 2 shows the amount of data for each subset and the number of phones for each language. It is worth noting that the amount of the transcribed data is only about 1 hour per language, which is significantly smaller than the usual amount of data used to train ASR systems.

	ENG	GER	HIN	JAP	MAN	SPA
Training [hours]	1.71	0.97	0.71	0.65	0.43	1.10
Validation [hours]	0.16	0.10	0.07	0.06	0.03	0.10
Test [hours]	0.42	0.24	0.17	0.15	0.11	0.26
Phone set size	39	43	46	29	45	38

Table 2. The OGI Stories corpus in terms of amount of data and number of phonemes used for each language.

3.1. Experimental Setup

A key goal for all the following cross-language experiments is to evaluate phone recognition performance on the target Japanese language by using the detection-based phone recognizer described in Figure 1. Two sets of experiments were conducted: (1) all detectors and mergers are trained without using Japanese-specific speech data; and (2) Japanese phone mergers are improved by incrementally adding Japanese data. We designed a series of experiments to analyze the performance of using different combinations of attribute detectors and event mergers. All ANNs were built with the ICSI QuickNet software package [10]. The Viterbi algorithm used to generate the recognized phone sequences was implemented with HTK [11]. To reduce acoustic mismatches across many recording conditions cepstral mean subtraction and unit variance normalization were applied on a per utterance basis for all speech data.

3.2. Without training data from the target language

3.2.1. Language-specific design

Here the attribute detectors and event mergers were trained with training data from a particular language [5]. For example, we use the notation, SPA-SPA, to denote the configuration that the attribute detector and event merger only use the one language-specific training set from Spanish. Therefore we report on five experimental configurations: ENG-ENG, GEM-GEM, HIN- HIN, MAN-MAN and SPA-SPA.

Table 3 lists performance of Japanese phone recognition in terms of phone accuracy rate (PAR). A uniform phone language model (0-gram) was used in all experiments. Among the five languages, the Spanish recognizer gave the best PAR. One possible explanation is that Spanish and Japanese share the same set of five vowels. We further analyzed the phone distribution of the test utterances in Japanese, and we found that the percentages were 45% and 55% for vowels and consonants, respectively. As a contrast for the English recognizer, the vowel resolution is too detailed than that for Japanese, because there are about 15 vowels in English. This can easily result in a poor performance shown in Table 3 with ENG-ENG. Another problem is that not all phones were defined in a similar way across all the six languages. As a result only a subset of Japanese phones can be trained for each set of language-specific attribute-to-phone mergers. For example there were only 26 phone mergers obtained in the SPA-SPA configuration resulting in phone recognition errors for the missing four phones and other units surrounding them in the Japanese test set. As for German we can only obtained a subset of 26.

Detector-Merger	ENG-ENG	GEM-GEM	HIN-HIN	MAN-MAN	SPA-SPA
PAR(%)	40.75	34.24	48.76	36.42	50.80

Table 3. The Japanese phone accuracy rates (PAR) using language-specific attribute detectors and phone mergers.

One way to reduce the level of difficulty of the missing phone definitions for a target language is to share phone mergers across different languages. We will address this issue later. Another solution is to better define phone sharing across different languages. This can come from some collaboration between speech researchers and speech processing engineers. This is one of the most serious issues to be investigated for research in multilingual ASR and universal phone recognition.

3.2.2. Attribute detector selection

According to the results of the Table 3, we observed that the Spanish and Hindi recognizers gave better performance on Japanese test utterances than the other three recognizers. After a careful examination, we also found that the detectors of these two languages also have the lower average detection error rates than the other three. In fact, these two languages are more suitable as the training languages to build a classifier to recognize Japanese sentences. Based on the model selection concept, we first selected the two languages to train the common set of 21 attribute detectors. We named this detector set as H+S. Then, based on the same selection concept, we build a set of detectors, named Select5, which selected and combined the detector with the highest attribute detection accuracy rate of the five languages. That means that the Select5 detectors attained the best attribute detection rate on Japanese test utterances.

One reason of conducting this experiment is to investigate if selecting detectors with better performances will also result in a better phone recognition accuracy rate on Japanese. The same scenario can be applied to selecting distinguished features. The more separable the features are the more reliable classifiers can be obtained. The results are shown in Table 4. Indeed, the performance of H+S-SAP is better than HIN-HIN, but when compared with SPA-SPA, the performance of H+S-SPA dropped a little possibly due to a mismatch set of Spanish mergers. On the other hand, the performance of Select5-SPA improved over all the systems in Table 3, including SPA+SPA. We believe a better selection strategy can further improve the system performance.

Detector-Merger	H+S-SPA	Select5-SPA
PAR(%)	50.20	52.49

Table 4. Japanese PARs on the selected attribute detectors.

3.2.3. Attribute-to-phone merger design

To investigate the potential of sharing data and attribute definitions among different languages while still maintaining good detection accuracies, we pool all the available training data from the other five languages and design a unified bank of detectors. We expend two languages in the H+S system in Table 4 to five languages, and named the detectors as 5L [5].

Besides the 5L-detectors, we tried two kinds of mergers: language-specific as defined before in Section 3.2.1, and phone-sharing (PG) to be discussed next. For the set of five languages we can find a collection of 211 unique phone symbols, some of them shared across different languages. We shared all 211 units and trained a corresponding collection of 211 attribute-to-phone mergers using all the speech data from all the five non-target languages. This phone merging scheme is called PS in the following.

Table 5 lists the phone accuracy rates of using 5L-detectors with language-specific and phone-sharing mergers. When compared with the language-specific results with those in Table 3, it indicates that additional improvements can be obtained by allowing the system to share data. This is a natural extension of the detector-based system because of the intrinsic universality of acoustic phonetic features. Next, when using the PS mergers, the PARs increased over language-specific mergers, which also means phone sharing helped with phone recognition. The relative improvement rate

was not so significant, but we believe further improvements from sharing phone definitions can be obtained by a refined definition of universal attribute-to-phone mapping.

Detector-Merger	5L-ENG	5L-GEM	5L-HIN	5L-MAN	5L-SPA	5L-PS
PAR(%)	45.96	36.82	49.00	44.54	52.50	53.04

Table 5. The Japanese phone accuracy rates of the 5-language attribute detector with language-specific, and PS mergers.

3.3. With training data from the target language

To analyze the effect of training data from the target language on the final performance, we added Japanese-specific speech data incrementally at four levels, 10%, 20%, 50% and 100%, to Spanish training data step by step to jointly train the phone mergers. Furthermore, we also trained a language-specific JAP-JAP system for comparison. As expected in the results shown in Table 6 we achieved better performance when more target-specific Japanese data was added to the Spanish data. The phone accuracy rates went from 53.35% to 55.33% as we added 10% to 100% of the Japanese training set. However, when we trained a Japanese-specific phone merger, the highest PAR of 61.83% was achieved. We believe this was a results of a mismatch in phone definition as observed in the H+S-SPA system in Table 4 when the detectors and mergers were not designed jointly in an optimal manner.

Detector-Merger	5L-[SPA+10% JAP]	5L-[SPA+20% JAP]	5L-[SPA+50% JAP]	5L-[SPA+100% JAP]	5L-JAP
PAR(%)	53.35	53.22	54.62	55.33	61.83

Table 6. Japanese phone accuracy rate when adding Japanese training data in the attribute detectors and/or event mergers.

4. Summary and Discussion

Starting with our pervious multi-language ASR study [5], we have extended our detection-based recognizer to phone recognizer design with little or no target-specific training speech data using various combinations of attribute detectors and phone mergers. We found that we achieved the best phone accuracy of Japanese phone recognition with no Japanese-specific training speech when using the 5L attribute detectors with phone-sharing phone merger. On the other hand, using the same 5L attribute detector and Japanese data to train the Japanese phone mergers, we achieved the best PAR of 61.83%. It is clear an area for improvement is in the area of attribute-to-phone merger design.

If we perform clustering on the set of 211 units obtained in Section 3.2.3 according some phonetic or acoustic definition we can obtain a smaller set of units to be shared across all languages. This phone merging scheme is called PC. Since we don't have a complete definition of a universal phone set, we can attempt to do phone clustering based on acoustic definitions. Suppose we have available a small set of target-specific validation data then we can generation some phone confusion matrices obtained when performing cross-language phone recognition using the systems defined in Section 3.2.1.

For example we generated confusion matrices of Spanish-Japanese and Hindi-Japanese recognition as shown in Figure 2. We can see that the phoneme, /ly/ in Spanish is more close to the Japanese phoneme /jh/ than that of the phoneme /jh/ in Spanish. Therefore, we can represent the original phone transcription of the OGI corpus with multiple phoneme candidates predicted by the cross-lingual confusion matrices and directly used them for phone mapping.

As shown with the probabilities in Figure 2, each Japanese phone has several phone candidates. For example for the Japanese vowel, /a_J/, the candidates were /aa_S/, /ah_E/, /a_H/, /aa_M/ and /A:_G/, obtained from the other five languages. After choosing an empirical threshold, we can discard the candidates with a probability value smaller than the threshold. This resulted in an average of five for each unseen Japanese phone to be used in PC merger design. This artificial experiment gave us the best PAR of 53.88% when we replace the PS mergers in Table 5 with the PC mergers. Clearly more research is needed.

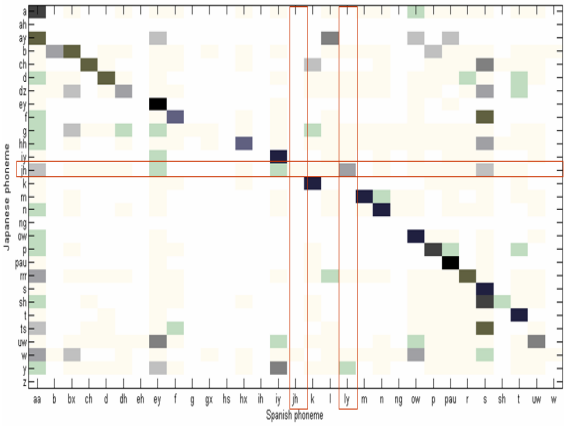

Figure. 2. An example of the confusion phoneme matrix.

5. References

[1] B. D. Walker, B. C. Lackey, J. S. Muller, P. J. Schone, "Language-reconfigurable universal phone recognition," Proc. of Eurospeech, 2003.

[2] J. Kohler, "Multilingual Phone Model for Vocabulary-Independent Speech Recognition Task," Speech Communication (35), 2001, pp. 21-30.

[3] International Phonetic Association, "Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet", Cambridge University Press, 1999.

[4] M. Siniscalchi, S., J. Li, and C.-H. Lee, "A study on lattice rescoring with knowledge scores for automatic speech recognition," in Interspeech, 2006.

[5] Sabato Marco Siniscalchi, Torbjorn Svendsen, and Chin-Hui Lee,"Toward A Detector-Based Universal Phone Recognizer," In Proc. of ICASSP, 2008.

[6] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "Towards bottom-up continuous phone recognition," in Proc. of ASRU'07, 2007.

[7] P. Mateˇjaka, P. Schwarz, J. Cˇ ernocky', and P. Chytil, "Phonotactic language identification using high quality phoneme recognition," in Proc. of Interspeech'05, 2005.

[8] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. of the IEEE, 77(2), Philadelphia, PA, 1989

[9] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," in Proc. of ICSLP' 92, 1992.

[10] http://www.icsi.berkeley.edu/Speech/qn.html

[11] S. Young et al., "The HTK Book", Version 3.2, 2002.