interspeech 2008 incorporating SST 08

Continuous Phone Recognition without Target Language Training Data

Dau-Cheng Lyu[1], *Sabato Marco Siniscalchi*[2] , *Tae-Yoon Kim*[3] *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, {tykim, chl}@ece.gatech.edu

Abstract

Designing an automatic speech recognition system with little or no language-specific training data is a challenging research topic because collecting abundant speech training data is not always an easy job for all possible languages of interest. According to our previous studied detection-based paradigm, we used a set of 21 acoustic phonetic attributes shared by five languages to perform Japanese phone recognition without using any Japanese speech training data. In this paper, we address the key issue of designing attribute-to-phone mapping models by two techniques: (1) a phone-based background model for each of the speech attribute detector to improve attribute detection; and (2) a data-driven clustering algorithm to group attribute-to-phone mapping rules of known languages to predict such rules for target phones in an unseen language. We report on experimental results of continuous Japanese phone recognition with the OGI Multilingual Speech Corpus and show that the proposed approach indeed decreases the false rejection rate of attribute detection, and improves the phone recognition accuracy.

Index Terms: unseen target language, speech attribute detection, detection-based speech recognition, phone cluster

1. Introduction

Due to fact that abundant resources in well-studied languages, such as English and Mandarin, are not always available for resource-limited languages, developing high performance automatic speech recognition (ASR) systems for these less-seen languages has always been an interesting research topic. Some studies addressed this issue as cross-language acoustic modeling [1, 2], and it can often be accomplished through: (1) bootstrapping [3, 4], and (2) language adaptation [1, 4]. The former uses acoustic models trained for other languages as seed models, and then language-specific training data is used for further refinement of the acoustic models. The latter is based on a well-trained acoustic model from a resource-abundant language to adapt to the target language using a small amount of language-specific speech data.

However, training a set of phone-based acoustic models for a target language without using any language-specific speech training data is still a challenging research issue. We can take advantage of the fact that many spoken languages share some sound patterns or even phones so that phone models for an unseen language may be predicted with this shared structure, e.g. through the usage of a large set of universal phones, such as the international phonetic alphabet (IPA) [5]. Another way is by defining a set of acoustic phonetic units, such as place and manner of articulation. Such attributes is believed to be more fundamental than phones and can be shared more extensively across a large number of languages [6].

Automatic speech attribute transcription (ASAT) [7] is a recently proposed paradigm aiming at developing ASR systems based on bottom-up detection of a set of speech attributes, such as voicing, nasality, dental, and frication, so that phone models can be built through attribute-to-phone mapping [8]. A possible implementation of an ASR system under the ASAT framework essentially integrates three levels of information [7]: (1) frame-based speech attribute detectors, (2) frame-based phone event mergers or attribute-to-phone mapping rules, and (3) decoding-based evidence verifiers.

In this paper, we expand upon our previous study on cross-language experiments [8] by further improving the attribute detectors and event mergers with phone-based background model (PBM) and data-driven phone clustering. First, the set of 21 detectors used in [7] in level 1 serves as a way to compute attribute posterior probabilities and can be trained with language-specific data, or with multilingual speech data from all five non-target (English, German, Hindi, Mandarin and Spanish) languages. In order to balance the amount of the training data between the attributes and non-attributes, a set of the non-attributes for each detector will be further divided into phone-based classes and the classes are regarded as the background models to compete with the attribute class. Secondly, for level 2, we used the five-language training data to build a five-language (5L) event merger with artificial neural networks (ANNs). Besides, the data-driven phone clustering mergers are used to cluster language-dependent phones. Finally, we used a Viterbi decoder as an evidence verifier to recognize the unseen target language (Japanese).

The rest of the paper is organized as follows. Section 2 describes a baseline system. Section 3 presents the proposed PBM-based attribute detectors. In Section 4, we show how to use data-driven methods to cluster phones in five non-target languages. In Section 5, continuous phone recognition results on unseen Japanese speech are presented and discussed.

2. Baseline System

As proposed earlier [6, 7] 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. The evidence verifier generates only the first best hypothesis as the recognized phone string for evaluating continuous phone recognition performance. Lattice rescoring has been shown to significantly improve phone and word recognition [9], and will not be addressed in the current study.

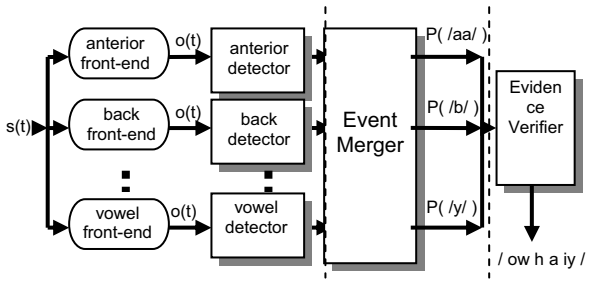

Figure. 1. Overall system

2.1. Speech Attribute Detectors

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 built each detector using 3 feed-forward ANNs with one hidden layer of 500 hidden nodes as organized in [8]. To estimate the ANN parameters, we separated 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 was used at the output layer to produce an approximate posterior probability that a particular speech event appears at the frame currently being processed. Energy trajectories in mel-frequency bands, organized in a split-temporal context as in [8], were used as 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 here were implemented using a single feed-forward ANN with one hidden layer of 800 hidden nodes. A softmax function was 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) [10]. 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 was performed over the decoding network to generate the decoded sequence of phones.

2.3. Speech Corpus and Experiment Setup

We adopted the "stories" part of the OGI Multi-language telephone speech corpus [11] for all experiments. The size of the training set of 5 languages, English, German, Hindi, Mandarin and Spanish, are 1.71, 0.97, 0.71 0.43 and 1.10 hours, respectively. The validation sets contain 0.16, 0.1, 0.07, 0.03 and 0.1 hours for the five languages. On the other hand, the unseen test language is Japanese, and the test set size is 0.15 hours. All of the six languages have available attribute and phonetic transcriptions. Each of the detectors contains three classes, positive (attribute), negative and others, and the numbers of the phones for the 5 training languages are 39, 43, 46, 44 and 38. The number of phones in Japanese is 29.

All ANNs were built with the ICSI QuickNet software package [12]. The Viterbi algorithm used to generate the recognized phone sequences was implemented with HTK [13]. 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. A uniform phone language model (0-gram) was used in all experiments.

2.4. Baseline Results

In Figure 2, we plot the phone accuracy rate (PAR) evaluating on the unseen Japanese utterances with language specific (LS) detectors and 5-language (5L) detectors. The former set was trained by language-specific training data, and the latter was a common bank of detectors which we pooled all the available training data from the other non-Japanese languages. Due to the potential of sharing attribute data, in general, the 5L detectors outperformed the LS detectors [8]. On the other hand, the Spanish recognizer gave the best PAR among the five languages and the PAR for LS and 5L detectors are 50% and 52.4%, respectively. One possible explanation is that Spanish and Japanese share the same set of five vowels.

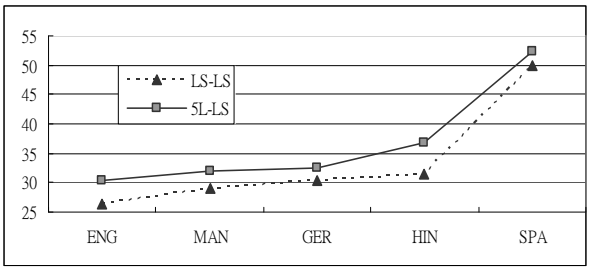

Figure 2. The baseline phone accuracy rate of using language specific (dotted line) and 5-language detectors (solid line).

2.5. Analysis

Clearly the results of the attribute detectors directly influence the performance of the merger and the final phone accuracy rate. We analyzed the attribute detectors and found that the detection accuracy for attributes, such as voiced and vowel, were over 90% if we used the evaluation measure in [14]. However, the performance for other attributes, such as glottal and approximant, were very low. We further found that the false rejection rates of the attribute class is much higher than that of the non-attribute class for glottal and approximant attributes. In the meantime the false acceptation rate of the non-attribute class was much higher than that of the attribute class. This implies that most of the test speech was recognized as the non-attribute class. Because of this consequence, we analyzed the data of the attributes of glottal and approximant, and found that an average occurrence of less than 5% of attribute segments was observed, which means that the other 95% data were used to only train the non-attribute classes. The condition of the data distribution is unbalanced, and the non-attribute class dominated the inputs to the ANNs. That is a reason of the high false rejection rate of the attribute class. In order to balance the distribution of the training data between attribute and non-attribute classes we used a phone-based background model to model the non-attribute classes as described in the following.

3. Phone-based Background Model (PBM)

One way to balance the data of negative and positive classes is to divide the non-attribute class into several subclasses. It resulted in a multiple-model representation of the non-attribute class. Although the total number of the classes increases in each detector, we only concern about the attribute, because only the outcome of the attribute class takes

up useful information. The other classes can be regarded as a background model, which consists of statistical models of an "average" negative class. This has been widely used as cohort modeling in speaker and utterance verification [15, 16].

In this study, we used information from phone category to cluster the background model, and we called it a phone-based background model (PBM). In other words, the training data of the non-attribute parts were segmented according to the phone categories. Take the approximant attribute in English as an example. The phones belonging to the approximant attribute are /w/, /y/, /l/ and /r/. The occurrence of the training data distributions of the approximant attribute is only 1%, and others are almost coming from the one non-attribute class. After using the PBM approach, the number of the non-attribute class increases to 37, and the average occurrence training data for each of the PBM is 2.6%. Therefore, the amount of the training data between the approximant attribute and the non-attribute classes are comparable. In Figure 3 (a) the histogram of posteriors obtained from the positive class samples in glottal detector is plotted. The heavy tail and peak at low posterior value indicate that the posterior estimate of the positive class is not reliable. It was clearly improved in Figure 3 (b) after PBM with more values peaked on the right.

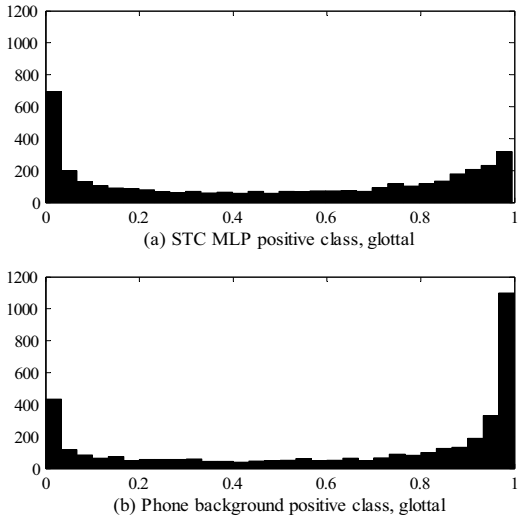

Figure 3. Histogram of posterior probabilities obtained from the MLP outputs in the positive class of the glottal detector

4. Phone Clustering for Event Merger

A promising result of using data sharing concept in Section 2.4 for attribute detection can be extended to phone merger design. In this following, we explore different approaches to phone clustering to predict attribute-to-phone mapping for unseen languages. The number of the Japanese phones, 29, is the smallest of the six languages. Several phones may exist in other languages that are similar to a given Japanese phone. Some form of language-specific or language-universal phone clustering may therefore benefit Japanese phone recognition.

We used a data-driven approach to merge different phones by measuring distance among phone models. For example Bhattacharyya distance [17] and a likelihood ratio test [18], have been used as confidence measures for phone distances. After clustering, phone models with similar characteristics are grouped in a broad phone class. The phones in the same group share the training data to train a single class model.

We used the posteriors of the attribute detectors as features to train a 3-state HMM for each phone. Each of state is modelled by a single Gaussian mixture model (GMM). For the similarity measure, we used the Bhattacharyya distance to measure distance between phone models as follows:

$$D_{pqi}=\frac{1}{8}(u_{pi}-u_{qi})^T\left[\frac{\Sigma_{pi}+\Sigma_{qi}}{2}\right]^{-1}(u_{pi}-u_{qi})+\frac{1}{2}\ln\frac{\left[\frac{\Sigma_{pi}+\Sigma_{qi}}{2}\right]}{\sqrt{|\Sigma_{pi}||\Sigma_{qi}|}}$$

where D_{pqi} is the Bhattacharyya distance between the i^{th} states of the p^{th} and q^{th} phones with u_{pi} and Σ_{pi} denoting the mean vector and covariance matrix of the i^{th} state in the p^{th} phone.

With all the distances computed between all phone pairs, we built a hierarchical agglomerative clustering (HAC) [19] tree. After that, a delta Bayesian information (△BIC) criterion [20] to merge language-dependent phones into a data-driven based language-independent phone in a bottom-up manner if the value of △BIC of the two phones (phone groups) is greater than zero. △BIC is defined as follows:

$$\Delta BIC=-\frac{n_p}{2}\log|\Sigma_p|-\frac{n_q}{2}\log|\Sigma_q|+\frac{n_r}{2}\log|\Sigma_r|+\frac{\lambda}{2}\left(d+\frac{d(d+1)}{2}\right)\log n_r$$

where n_p, n_q and n_r are the number of occurrences at node *p, q* and *r*; Σ_p, Σ_q and Σ_q are the covariance matrices of model *p, q* and *r*, respectively. The λ value we used here is 1, and *d* is the number of feature vector dimensions of the model.

The first three terms of above equation signify the class similarities due to the differences between the individual and grouped classes. The last term is a penalty term in proportion to the model complexity and the number of grouped training frames. Therefore, using △BIC as the criterion to cluster the phone models offers two benefits. One is to use an optimal Bayesian model-selection criterion to measure the difference, and the other is to consider a balance between the model complexity and the amount of the training data.

This procedure is similar to a tandem speech recognition system [21], which is to use ANN outputs as features to train phone targets and then to generate input features that feed into a conventional GMM-based recognizer. However, in our task, we used ANNs to train attributes and then utilized the posteriors as features to train the GMM. Besides, the GMMs were trained in order to cluster the phones if they are similar, and the final mergers still use ANNs.

We take the vowels of German as an example, and show the phone clustering result in Figure 4. Duo the use of △BIC, we grouped the phones and each blocked unit was regarded as a new acoustic unit. Then, we used 0.06 hours of Japanese speech as a cross validation dataset to test the newly trained German merger, and found the mapping rule between the acoustic units and the Japanese phones for recognition.

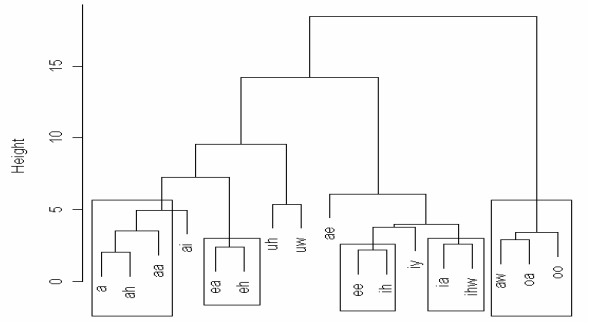

Figure 4. German vowels before and after phone clustering.

5. Experiment Results

The results on Japanese data with language-specific and 5L attribute detector using the PBM approach are shown in figure 5. In general, we observed an average improvement of 1.9% and 1.8% over the results showed in figure 5 for the language specific and the 5L attribute detectors, respectively.

Furthermore, the accuracy of 5L-LS-PC, using data-driven clustering, outperforms that of 5L-LS. Though the accuracy of 5L-LS-PC in Spanish was almost the same as that of 5L-LS, improvements have been observed for other languages as well. The reasons are caused by the less number of phone candidates during the decoding. In fact, the numbers of acoustic units for the five languages reduced to 31, 29, 33, 32, and 26 of ENG, MAN, GER HIN and SAP after the phone clustering step.

Besides, based on the good performance of 5L-LS-PC, we selected the acoustic models from the five non-target languages. The models better represented the Japanese phones. For example, the 5 Japanese vowels are covered by /aa/, /ey/, /iy/ and /ow/ from SPA and /uw/ from HIN. The selected acoustic models are implemented within a single merger. This merger consisted of 26 acoustic models from 5-non-target languages, and they are 9 phones from SPA, 8 from HIN, 5 from ENG and 4 from GER. After testing on the Japanese test data, we have achieved 55% phone accuracy rate which is the highest performance in this paper.

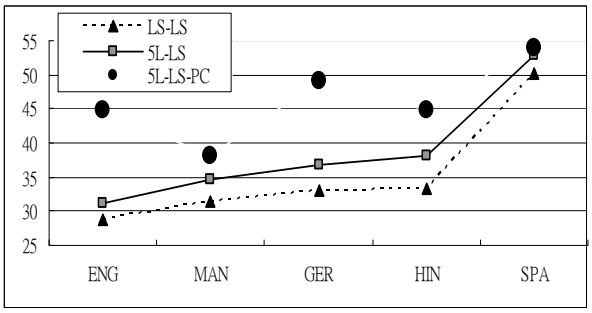

Figure 5. PBM phone accuracy of using language specific detector (dotted line) and 5-language detector (solid line). The circle points are the results of using phone clustering.

6. Summary and Future Work

We have extended our detection-based recognition system to phone recognition of unseen languages with no training data from the target language [6]. Firstly, with phone background modeling in detector design we not only improved attribute detection, but also increased the phone accuracies. Secondly, we found that we achieved the best phone accuracy of Japanese phone recognition using the 5L attribute detectors with data-driven phone clustering to build event mergers for unseen Japanese phones. Attribute-to-phone mapping rules are typically very difficult to design with no language-specific training data. We will be exploring other ideas in future studies. In the meantime, we believe a new set of universal phone units taking into account of linguistic and acoustic definitions will facilitate a effective prediction of mapping rules for phones not previous seen in new or resource-limited languages. This will definitely enhance the capabilities and performance of universal phone modelling and recognition.

7. Acknowledgment

We greatly appreciate Taiwan's National Science Council and Korea's Research Foundation Grant KRF-2007-357-D00195 for partially supporting the first and third authors for their visiting stay at Georgia Tech.

8. References

[1] A. Constantinescu and G. Chollet, "On cross-language experiments and datat-driven units for alisp," Proc. ASRU, 1997.

[2] T. Schultz and A. Waibel, "Experiments on cross-language acoustic modeling," Proc. Eurospeech, 2001.

[3] A. Wheatley, K. Kondo, W. Anderson, and Y. Muthusamy, "An evaluation of cross-language adaptation for rapid hmm development in a new language," Proc. ICASSP, 1994.

[4] T. Schultz and A. Waibel, "Language independent and language adaptive lvcsr," Proc. ICSLP, 1998.

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

[6] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "Towards bottom-up continuous phone recognition," Proc. ASRU, 2007.

[7] C.-H. Lee, et al. "An overview on automatic speech attribute transcription (ASAT)" Proc. Interspeech, 2007.

[8] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee,"Toward A Detector-Based Universal Phone Recognizer," Proc. ICASSP, 2008.

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

[10] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, 77(2), 1989.

[11] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," Proc. ICSLP, 1992.

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

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

[14] J. Li and C. -H. Lee, "On designing and evaluating speech event detectors," Proc. Interspeech, 2005.

[15] A. E. Rosenberg, et al, "The Use of Cohort Normalized Scores for Speaker Verification," Proc. ICSLP, 1992.

[16] R. A. Sukkar and C.-H. Lee, "Vocabulary Independent Discriminative Utterance Verification for Non-Keyword Rejection in Subword Based Speech Recognition," IEEE Trans. on Speech and Audio Proc., 4(6), 1996.

[17] B. Mak and E. Barnard, "Phone Clustering Using the Bhattacharyya Distance," Proc. ICSLP, 1996.

[18] Liu Yi and Pascale Fung, "Automatic Phone Set Extension with Confidence Measure for Spontaneous Speech," in Proc. of EuroSpeech 2005.

[19] T.S. Chen, C.C. Lin, Y.H. Chiu and R.C. Chen "Combined Density-and Constraint-based Algorithm for Clustering," In Proc. of 2006 ICISKE

[20] G. Schwarz, "Estimating the dimension of a model," The annals of statistics, vol. 6, 1978, pp 461-464

[21] H. Hermansky, D. P. W. Ellis and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," Proc. ICASSP, Istanbul, Turkey, 2000.