TOWARD A DETECTOR-BASED UNIVERSAL PHONE RECOGNIZER

Sabato Marco Siniscalchi[1], Torbjørn Svendsen[1], and Chin-Hui Lee[2]

[1]Department of Electronics and Telecommunications
Norwegian University of Science and Technology, Trondheim, Norway
[2]School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA 30332 USA
{marco77, torbjorn}@iet.ntnu.no, chl@ece.gatech.edu

ABSTRACT

In recent research, we have proposed a high-accuracy bottom-up detection-based paradigm for continuous phone speech recognition. The key component of our system was a bank of articulatory detectors each of which computes a score describing an activation level of the specified speech phonetic features that the current frame exhibits. In this work, we present our first attempt at designing a universal phone recognizer using the detection-based approach. We show that our technique is intrinsically language independent since reliable articulatory detectors can be designed for diverse languages, and robust detection can be performed across languages. Moreover, a universal set of detectors is designed by sharing the training material available for several diverse languages. We further demonstrate that our approach makes it possible to decode new target languages by neither retraining nor applying acoustic adaptation techniques. We report phone recognition performance that compares favorably with the best results known by the authors on the OGI Multi-language Telephone Speech corpus.

Index Terms— Detectors, speech recognition, knowledge based systems

1. INTRODUCTION

Over the past few years, several researchers have been challenged to design a universal ASR system for multiple languages [1, 2, 3, 4, 5, 6]. The benefits associated with a universal ASR systems are several, for example: 1) unseen target languages can be decoded, 2) the number of model parameters can be reduced, and 3) training material can be shared. The design of a unified speech recognizer is not simple, as it requires defining a *universal phonetic alphabet* (UPA) along with a mechanism to handle language-dependent variations.

The design of a UPA has been one of the topics explored at the 2007 Jonh Hopkins University Summer Workshop hosted by the Center for Language and Speech Processing. Yet, it is still far from being completed. Thus, a unified phonetic inventory is usually generated for the languages at hand by using phoneme mapping techniques either guided by acoustic-phonetic knowledge, for example using the International Phonetic Alphabet (IPA) [7], or by data-driven techniques which find acoustic similarities across sounds of several languages, e.g., [1, 2, 3, 4, 5]. In [6], a bottom-up, two-level forced alignment technique is used to generate consistent phonetic labeling across several corpora.

It is a common practice to address the problem of decoding a new target language as an acoustic modeling problem, e.g., [8, 4, 5]. Language-independent acoustic modeling paradigms can be often accomplished through: 1) *bootstrapping* [9, 10], and 2) *language adaptation*, e.g., [10, 4]. The main idea of bootstrapping is to use acoustic models trained for other languages as seed models for a new target language; then language specific training data is used for further refinement of the acoustic models. Language adaptation techniques follow the idea underlining the speaker adaptation techniques. Thus, acoustic models built for other languages are adapted to the target language using small amounts of data. Finally, cross-language experiments refer to experimental setups in which the target language material has not been used during the training phase, e.g. [8, 5, 11]

In this paper, we describe our first attempt at designing a universal phone recognizer (UPR) which can decode a new target language with neither adaptation nor retraining. We mainly focus on acoustic modeling problems. We believe that the first step toward building a UPR is to identify a common knowledge source (KS) that is fundamental and sharable across languages. We propose articulatory features as KS since sounds described by the same set of these features are similar across languages. The second step is to include the linguistic knowledge carried by the articulatory features into the UPR design process. In [10, 5], linguistic information has been incorporated while modeling phones in context, yet little improvement was reported in cross-language experiments. The authors argued this was due to the poor context overlap of different languages. In [12], we incorporated articulatory information directly into acoustic modeling of context-independent phone models. We extracted articulatory motivated features by a bank of detectors and used these features as basic units for acoustic modeling. In the present work, we extend our technique to a multilingual scenario provided by the OGI Multi-language Telephone Speech corpus [13]. We show that it is possible to detect articulatory motivated features reliably for different languages, and robustly across languages. Moreover, we show that parameter reduction, with no loss in phone recognition performance, can be achieved by using a universal bank of detectors. The universal bank of articulatory detectors is built by sharing language specific data. Finally, we give insights into the cross-language capability of the proposed approach by training on other languages and decoding a never seen target language.

The topic of articulatory features in a multilingual context was previously explored in [14]. Our work differs from that work in several aspects. First, our system makes sole use of articulatory motivated features. Second, we define manner and place of articulation for consonants and vowels into a common linguistic space as suggested by [15]. Finally, in cross-language experiments, we evaluate our UPR system on a never seen target language.

This wok is partially supported by SIRKUS project. The authors thank Daucheng Lyu, a Ph.D. student visiting Georgia Tech from Chang Gung University in Tao-Yuan, Taiwan.

2. ARTICULATORY ATTRIBUTES

In this work, we assume that sounds described by the same set of articulatory features share common acoustic properties across languages. Consequently, articulatory features can be considered as more fundamental units than phonemes, since they are independent of the underlying language. As pointed out in [15], a difficulty when using manner and place of articulation for ASR application is that vowels and consonants cannot be mapped into a common linguistic space because place of articulation has been differently defined for them. We follow [15], and we force vowels and diphthongs to be organized into the place classes as the consonants. Also, we consider all articulatory features as binary although some of them take on non-binary discrete and continuous values. Furthermore, we increase the set of manner and place categories to provide some redundancy by some of the distinctive features defined by Chomsky and Halle [16]. We use English as a rule model to define the mapping from phonemes to distinctive features for all the other languages. In this work, we use the terms attribute to refer to the set of articulatory features used. Table 1 shows the set of attributes that we use in our experiments along with the attribute-to-phone for English. Some of the phone-to-attribute mappings may be arguable, but they are based on common practice in speech recognition literature.

3. SYSTEM ARCHITECTURE

Figure 1 shows a block diagram of our detector-based UPR system which consists of three main blocks: (1) a bank of speech event detectors, (2) an event merger, and (3) an evidence verifier. More details about each block will be provided in the following sections. The evidence verifier generates only the first best hypotheses.

3.1. Speech Event Detectors

The goal of each detector is to analyze the speech signal and produce a confidence score or a posterior probability that pertains to some acoustic-phonetic attribute. We build each detector using 3 feed-forward ANNs with one hidden layer and 500 hidden nodes organized as in [12]. To estimate the ANN parameters , we separate the training data into attribute present and attribute absent regions for every articulatory event using the available phonetic transcription. The softmax activation function is used in the output layer, and the ANN produces the posterior probability that a speech event has happened during the frame being processed. Energy trajectories in mel-frequency bands which are organized in split-temoporal context as in [17] are used as parametric representations of the speech signal.

3.2. Event Merger & Evidence Verifier

The event merger combines the event detectors' outputs together with different weights and delivers evidences at a phone level. The event merger is implemented using a single feed-forward ANN with one hidden layer and 800 hidden nodes. The softmax activation function is used in 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 probability. Each phone is modeled by a 3-state left-to-right hidden Markov model (HMM). The HMM state likelihood is the phone posterior probability of the event merger. We assume equal prior probabilities for all phones. The Viterbi algorithm performed over the decoding network provides the decoded sentence.

4. EXPERIMENTAL SETUP

All the experiments were conducted using the "stories" part of the OGI Multi-language telephone speech corpus [13] which has phonetic transcription for six languages: English (ENG), German (GEM), Hindi (HIN), Japanese (JAP), Mandarin (MAN), and Spanish (SPA). For each language, we divided the database into three subsets, namely: training, validation, and test. Table 2 shows the amount of data for each of these subsets and the number of phones for each language. It is worth to point out 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.

All the ANNs were built using the ICSI QuickNet neural network software package[1], and are trained using the classical back-propagation algorithm with cross entropy error function. The Viterbi algorithm used to generate the recognized phone sequence was implemented using the HTK toolkit[2].

Attribute	Phoneme set
Vowel	iy ih eh ey ae aa aw ay ah oy ow uh uw er
Fricative	jh ch s sh z f th v dh hh
Nasal	m n ng
Stop	b d g p t k dx
Approximant	w y l r er
Coronal	d dx l n s t z
High	ch ih iy jh sh uh uw y ey ow g k ng
Dental	dh th
Glottal	hh
Labial	b f m p v w
Low	aa ae aw ay oy ah eh
Mid	ah eh ey ow
Retroflex	er r
Velar	g k ng
Anterior	b d dh dx f l m n p s t th v z w
Back	ay aa ah aw ow oy uh uw g k
Continuant	aa ae ah aw ay dh eh er r ey l f ih iy oy ow s sh th uh uw v w y z
Round	aw ow uw uh v y oy r w
Tense	aa ae aw ay ey iy ow oy uw ch s sh f th p t k hh
Voiced	aa ae ah aw ay b d dh dx eh er ey g ih iy jh l m n ng ow oy r uh uw v w y z
Silence	pauses

Table 1. *American-English phonemes list in terms of the manner and place of the articulation.*

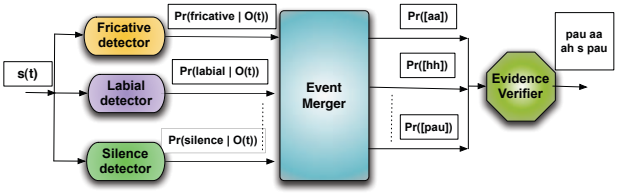

Fig. 1. Detector-based phone recognition system

4.1. Attribute Detection Experiments

4.1.1. Language Dependent Attribute Detection

Each detector is trained, validated, and tested using only language specific data, as shown in Table 2. Moreover, each detector classifies a given speech frame as either attribute present or attribute

[1] ICSI quicknet software package, http://www.icsi.berkeley.edu/speech/qn.htm

[2] HTK toolkit, http://htk.eng.cam.ac.uk/

Language	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
Phoneme set	40	44	47	30	45	39

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

Attribute	ENG	GER	HIN	JAP	MAN	SPA
back	77.12	67.71	85.69	87.53	88.08	76.24
cont.	81.01	73.76	79.41	88.82	89.49	88.43
fric.	75.83	83.06	73.48	75.74	80.02	71.88
glot.	37.57	38.07	31.84	33.33	42.06	41.19
approx.	62.08	44.87	45.73	40.74	48.40	55.39
high	70.88	65.82	73.75	71.13	75.52	61.34
labial	68.50	64.26	68.64	54.24	41.53	69.33
mid	71.48	74.58	77.82	83.41	81.33	80.88
nasal	68.87	77.70	67.69	69.30	62.86	74.48
retr.	61.81	46.22	47.12	46.02	52.86	46.75
tense	78.37	81.99	85.95	93.12	79.35	91.46
voiced	88.44	91.00	90.00	94.12	90.09	92.67
vowel	91.70	94.73	91.62	93.12	92.18	93.09

Table 3. *Language-specific accuracy rates (%) for several speech attributes.*

absent. Table 3 shows the accuracy rates for each language and for several attributes[3]. First, attribute accuracies are comparable across languages and attributes. We achieve reliable attribute accuracy for diverse languages, for example, the attribute accuracy is as high as 92% on the vowel class. For the fricative class, we achieve lower accuracy as compared to common rates reported on wide-band, but this is due mainly to 4 kHz cut-off frequency of telephone speech. Despite the more available training data for English, there are cases where other languages achieve a better attribute accuracy. For example, vowels are better detected in Spanish than in English. The main explanation is that the English vowel class is larger than the Spanish one in terms of the number of phonemes. Finally, from a qualitative point of view, our results are comparable with Stüker's [14], but we use telephone quality speech material, much less training data, and we do not restrict training and evaluation to the middle part of the speech event.

4.1.2. Cross-language and Universal Attribute Detection

The intent of this section is twofold: 1) we want to study whether robust detection can be carried out across languages, and 2) we want to investigate the possibility of sharing data among different languages, and thereby the possibility of designing a single bank of detectors for several and diverse languages. To address the robustness issue, we have tested all the detectors of a specific language on the data of the other languages. Due to space constraints, we only report cross-language experiments when Mandarin (MAN) is used as a test set. Figure 2 show accuracy rates for each single detectors and language on MAN data. The connected line in the figure shows the attribute accuracy of Mandarin detectors on the MAN test data. Accuracy rates for several speech attributes across languages are less reliable than in the language-dependent case, yet the drop in performance is not particularly severe. For example, the attribute accuracy for the

[3]The glottal attribute achieves the lowest accuracy rate among all of the six languages.

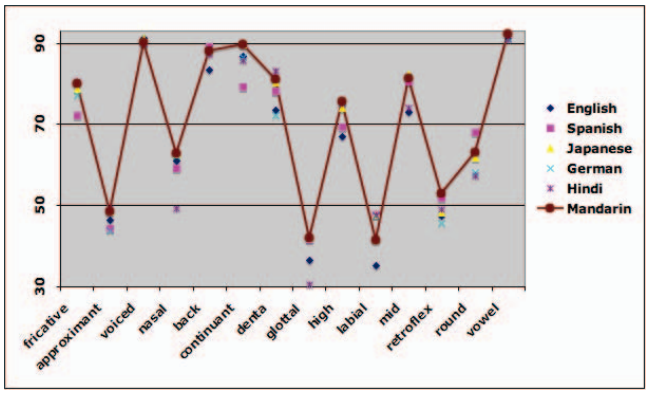

Fig. 2. Cross-detection accuracy using Mandarin as test set. The connected line shows the attribute accuracy of Mandarin detectors on the Mandarin test set

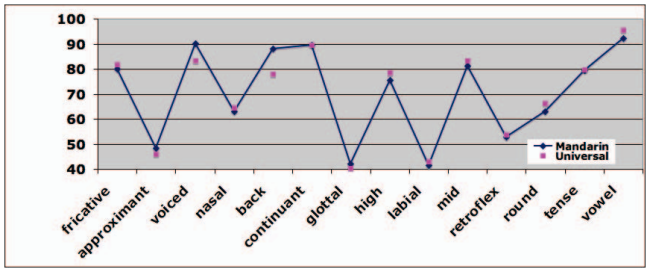

Fig. 3. Universal detectors accuracy on Mandarin test set.

nasal class are comparable across all languages. Furthermore, for some attributes, such as vowel, the data points are either on or above the connected line. This indicates that the vowel detector trained on a language other than Mandarin achieves an attribute accuracy comparable with the corresponding Mandarin detector one. Thus, detectors from different languages may be selected to achieve the overall best performance for each attribute. To investigate the possibility of sharing data and parameters among different languages while still achieving good attribute accuracies, we pool all the available training data from all the six languages and design a unified bank of detectors (*Universal*). Figure 3 shows attribute accuracies on Mandarin test sentences for several classes. We can observe that by pooling the data, better attribute accuracies are achieved for several attributes; for example, vowel, fricative, and mid.

4.2. Phone Recognition Experiments

4.2.1. Language Specific Phoneme Recognition Experiments

The language independent experiments discussed earlier reveal that it should be possible to train a single bank of attribute detectors for all the languages to build our detector-based phone recognizer. To show this, we performed two sets of experiments. First, six individual recognizers were designed by training both the bank of detectors and the evidence merger on language specific data (L-Specific). In the second experiment, Universal bank of detectors was used, and the evidence merger was still trained on language specific data (L-Universal). Table 4 shows the performance, in terms of phone error rate (PER), for the two configurations. A 0-gram language model is used in all experiments. To the authors' knowledge, the BUT systems [17] reports the best results on the same data sets. The BUT performance is shown in the bottom row of the table.

When detectors are trained on language specific data, our sys-

Language	ENG	GER	HIN	JAP	MAN	SPA
L-Specific	46.68	50.84	45.55	39.95	49.19	39.99
L-Universal	45.24	49.93	43.03	38.22	47.42	38.75
BUT	45.26	46.10	45.74	41.19	49.93	39.55

Table 4. *PERs, in percentage (%), on the OGI Stories test sentences. The last rows refers to BUT performance as reported in [17].*

tem achieves comparable phone error rates (PERs) with the BUT system for all the languages, but German. In [12], we have already shown that a better performance can be achieved by increasing the number of detectors. The second row of Table 4 indicates that additional improvements can be obtained by allowing the system to share data and parameters as well. Data and model sharing is a natural extension of our detector-based system, for the intrinsic universality of the articulatory features. Moreover, our L-Universal phone recognizer outperforms the BUT system for all the languages, except for German. We think the German result comes from a possible incorrect phoneme-to-attribute mapping. Indeed, we used English attribute-to-phone mappings reported in Table 1 as a role model to accomplish this mapping, rather then expert phonetician knowledge. We believe that improvement on the German task can be obtained by refinement of the aforementioned mapping.

4.2.2. Cross-language Phoneme Recognition Experiments

For cross-language experiments, we designed a Spanish recognizer (DT-SYS) by training the detectors on all the available data, except for the Japanese. We also built a "traditional" MFCC based context-independent HMM phone recognizer, trained on Spanish material, which serve as baseline. The MFCC features are a thirteen component cepstrum vectors concatenated with the first and the second difference cepstrum. Continuous diagonal mixture Gaussian observation density HMM models are used. Evaluation was performed on never seen before Japanese test sentences. A unified universal phone set was built for Spanish and Japanese using knowledge-based rules. A 0-gram language model was used. Table 5 shows the PERs. Baseline's performance is given for different number of mixtures. We can draw several conclusions from Table 5. First, the DT-SYS outperforms the baseline system in all cases. The improvement in baseline's performance achieves a plateau at thirty-two mixtures, and consequently DT-SYS's better performance is not due to a higher system's complexity. We can also observe that although lower PER is achieved by the DT-SYS system when language specific material is used Table 4, the performance drop is far less than what could be expected when no language specific data is used. We believe that better results can be obtained by refining the attribute accuracies, for those accuracies are affected by the quality of the attribute-to-phone mapping tables.

5. CONCLUSION

We have extended our detection-based paradigm [12] to a multi-language scenario, and we have reported experiments for all languages both at an attribute and a phone level. We have shown that our detection based system always achieves phone accuracy results that are comparable or better than the, to the author's knowledge, best reported results OGI Multi-language Telephone Speech corpus. We have also presented our preliminary attempts to design a universal phone recognizer which is a viable solution for processing speech in all languages, even on those when no training data is available at all. The fact that this system could be built on one set of languages and tested on a never seen language is a clear indication of its potential.

System	DT-SYS	baseline 16-mix	baseline 32-mix	baseline 40-mix	baseline 44-mix
PER (in %)	47.5	54.2	52.9	52.5	52.4

Table 5. *Cross-language performance on Japanese sentences, as percentage of PER.*

6. REFERENCES

[1] J. Köhler, "Multilingual phoneme recognition exploiting acoustic-phonetic similarities of sounds," in *Proc. of ICSLP'96*, 1996.

[2] P. Cohen, S. Dharanipragada, J. Gros, M. Monkowski, C Neti, S. Roukos, and T. Ward, "Towards a universal speech recognizer for multiple languages," in *Proc. of ASRU'97*, 1997.

[3] S. Goksen and Gokcen J. N., "A multilingual phoneme and model set: towards a universal base for automatic speech recognition," in *Proc. of ASRU'97*, 1997.

[4] W. Byrne, P. Beyerlein, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, D. Vergyri, and W. Wang, "Towards language independent acoustic modeling," in *Proc. of ICASSP00*, 2000.

[5] T. Schultz and A. Waibel, "Experiments on cross-language acoustic modeling," in *Proc. of Eurospeech'01*, 2001.

[6] B. D. Walker, B. C. Lackey, and P. J. Muller, J. S.and Schone, "Language-reconfigurable universal phone recognition," in *Proc. of Eurospeech'03*, 2003.

[7] A.B. Smith, C.D. Jones, and E.F. Roberts, "Handbook of the international phonetic association," *Cambridge University Press*, 1999.

[8] A. Constantinescu and G. Chollet, "On cross-language experiments and datat-driven units for alisp," in *Proc. of ASRU'97*, 1997.

[9] A. Wheatley, K. Kondo, W. Anderson, and Y. Muthusamy, "An evaluation of cross-language adaptation for rapid hmm development in a new language," in *Proc. of ICASSP'94*, 1994.

[10] T. Schultz and A. Waibel, "Language independent and language adaptive lvcsr," in *Proc. of ICSLP'98*, 1998.

[11] K. Shinoda and C.-H. Lee, "A structural bayes approach to speaker adaptation," *IEEE Trans. Speech Audio Processing*, vol. 9, Nov. 2001.

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

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

[14] S. Stüker, T. Schultz, F. Metze, and A. Waibel, "Multilingual articulatory features," in *Proc. of ICASSP'03*, 2003.

[15] M. Tang, S. Seneff, and V. W. Zue, "Modeling linguistic features in speech recognition," in *Proc. of Eurospeech'03*, 2003.

[16] N. Chomsky and M. Halle, *The Sound Pattern of English*, MIT Press, 1991.

[17] P. Matějaka, P. Schwarz, J. Černocký, and P. Chytil, "Phonotactic language identification using high quality phoneme recognition," in *Proc. of Interspeech'05*, 2005.