Articulatory-Acoustic-Feature-based Automatic Language Identification

Supphanat Kanokphara and Julie Carson-Berndsen

School of Computer Science and Informatics
University College Dublin, Belfield, Dublin 4, IRELAND
{supphanat.kanokphara, julie.berndsen}@ucd.ie

Abstract

Automatic language identification is one of the important topics in multilingual speech technology. Ideal language identification systems should be able to classify the language of speech utterances within a specific time before further processing by language-dependent speech recognition systems or monolingual listeners begins. Currently the best language identification systems are based on HMM-based speech recognition systems. However, with the cost of this low percentage error, comes an increase in computational complexity. This paper proposes an alternative way of using HMM-based speech recognition systems. Instead of using phoneme level acoustic models and n-gram language models, articulatory feature level acoustic models and n-gram language models are introduced. With this approach, the computational complexities of language identification systems are considerably reduced due to the fact that the size of the articulatory feature inventory is naturally smaller than that of the of phoneme inventory.

1. Introduction

A truly multilingual speech recognition system is one of the dreams in many international businesses. For example, in a hotel where customers come from many different countries, it would be nice if a machine could be provided capable of communicating with customers in their own language. The system should be able to identify the language and recognize the speech. Practically speaking, identifying the language during recognition would require many speech recognizers (one for each language) running in parallel which prohibits real time applications. One way to solve this problem is to run only a language identification system and let this system choose the most likely language-dependent speech recognition systems which can then be used to recognize the speech.

While many approaches to language identification systems have been investigated such as spectral-similarity, prosody etc. the HMM-based speech recognition has proved to be the best for language identification thus far due to high-level knowledge in the systems [1]. HMMs are predominantly used as acoustic models in most speech recognition systems. This is because speech signal is varied differently in both time and signal amplitudes and HMMs are designed to cope with this kind of signal. The hidden state property in HMMs can normalize the time-variation while the statistic parameters in each state can cover the signal amplitudes. The statistic parameters and the transitions in HMMs are usually trained and optimized from several speech examples. In speech recognition, each acoustic model is generally used to represent each phoneme in speech. However, if the acoustic model is representing a larger unit like the whole utterance or even each word in the utterance, the models become too large to build. Normally, in speech recognition systems, a sequence of phonemes is mapped to a speech signal under constraints from language-specific dictionary and word-level n-gram language model.

The articulatory-acoustic features (AF, also called articulatory features or phonological features by other authors) are smaller units than phonemes that can be represented by acoustic models. Therefore, the system accuracy from AF models is commonly worse than the accuracy from phoneme models due to the fact that the units are smaller and contain less acoustic information. However, this has several advantages in a multilingual environment. Firstly, AFs share acoustic information from many phones. This makes AF models more robust to multilingual noise than their phoneme counterparts. Secondly, since AFs are typically similar in many languages, portability from one language to another becomes easier. Thirdly, the smaller number of models required by the AF approach, the shorter the processing time.

This paper presents an alternative way of using HMM-based speech recognition systems for the purposes of language identification. Instead of using phoneme models, AF models are introduced. Since, in general, language identification accuracy is relatively high, compared with the speech recognition accuracy, the fact that there is may be a degradation in performance in speech recognition accuracy using AF models, the AFHMM-based language identification systems should still be acceptable while the system complexity is reduced considerably.

The rest paper is organized systematically as follows. Section 2 describes the corpora used and the language identification systems which were built. Section 3 describes the experiments and presents the experimental results while section 4 draws the conclusions and briefly sketches some future work.

2. System Description

2.1. The Corpora

Two corpora are used in this paper (TIMIT for English [2] and NECTEC-ATR for Thai [3]). The TIMIT corpus consists of 3,600 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the U.S., of which 462 are in training set and 168 are in the test set. There is no overlap between the training are test sentences, except 2 dialect (SA) sentences which were read by all speakers. The training set contains 4,620 utterances and the test set 1,680 (112 males and 56 females). The core test set, which is the

abridged version of the complete test set, consists of 192 utterances (no SA sentences), 8 from each of 24 speakers (2 males and 1 female from each dialect region). SA sentences are eliminated from the language model training process but still exist in acoustic model training process. SA sentences are removed because they occur in both training and test set. The core test set is used for testing.

For NECTEC-ATR corpus, the speech database consists of 390 Thai phonetically balanced (PB) sentences. The vocabulary size is 1,476 words. The average number of words per sentence is 10. The average number of phones per word is 3.6. 42 speakers (21 males and 21 females) are separated into 34 speakers (17 males and 17 females) for training and 8 speakers (4 males and 4 females) for testing. Speakers for training are required to read 376 from 390 sentences while speakers for testing read another 14 sentences. All utterances are recorded in an office environment.

2.2. The Feature Table

In order to train the AF models, each phoneme in each of the corpora is transformed into corresponding AF. AFs can be classified according to properties which have mutually exclusive values. Each AF property is known as a *tier* in a multilinear representation. For example, *voiced* and *unvoiced* features can be classified as on a *voice tier*. In this paper, 5 *tiers* are evaluated (*voice*, *manner*, *place*, *height* and *type*). Table 1 shows the AFs used for TIMIT phonemes. The feature table for Thai can be found in [4].

For the language identification process, the word-to-phoneme lexicons from both corpora are converted to word-to-AF lexicons according to these feature tables.

2.3. HMM-Based Systems

The language identification systems in this paper have been constructed using HTK [5] which is now widely used for HMM-based speech recognition experiments. The acoustic model training process starts by converting the speech signal into a sequence of vector parameters with a fixed 25 ms frame and a frame rate of 10 ms. Each parameter is then pre-emphasized with the filter $P(z) = 1-0.9*z^{(-1)}$. The discontinuities at the frame edges are attenuated by using Hamming window. A fast Fourier transform is used to convert time domain frames into frequency domain spectra. These spectra are averaged into 24 triangular bins arranged at equal mel-frequency intervals (where $f_{mel} = 2595log10(1+f/700)$). f denotes frequency in *Hz*. 12 dimensional mel-frequency cepstral coefficients (MFCCs) are then obtained from cosine transformation and lifter. The normalized log energy is also added as the 13th front-end parameter. The actual acoustic energy in each frame is calculated and the maximum found. All log energies are then normalized with respect to maximum

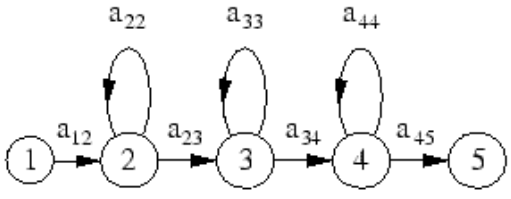

Figure 1: 5-state left-right HMM.

Table 1: The feature table for TIMIT phonemes

phonemes	voice	manner	place	height	type
aa	voiced	voc	cen	low	ten
ae	voiced	voc	frt	low	ten
ah	voiced	voc	cen	semilo	ten
ao	voiced	voc	bak	semilo	ten
aw	voiced	voc	cen	nil	ten
ax	voiced	voc	cen	mid	lax
ax_h	voiced	voc	cen	mid	lax
axr	voiced	voc	ret	nil	lax
ay	voiced	voc	cen	nil	ten
b	voiced	stp	lab	nil	nil
bcl	voiced	stp	lab	nil	nil
ch	unvoiced	frc	palv	nil	nil
d	voiced	stp	alv	nil	nil
dcl	voiced	stp	cor	nil	nil
dh	voiced	frc	den	nil	nil
dx	voiced	flp	alv	nil	nil
eh	voiced	voc	frt	semilo	lax
el	voiced	latapp	alv	mid	lax
em	voiced	nas	lab	nil	nil
en	voiced	nas	alv	nil	nil
eng	voiced	nas	vel	nil	nil
epi	sil	sil	sil	sil	sil
er	voiced	voc	ret	nil	lax
ey	voiced	voc	frt	nil	ten
f	unvoiced	frc	lab	nil	nil
g	voiced	stp	vel	nil	nil
gcl	voiced	stp	vel	nil	nil
h#	sil	sil	sil	sil	sil
hh	unvoiced	frc	glo	nil	nil
hv	voiced	voc	cen	nil	lax
ih	voiced	voc	frt	semihi	lax
ix	voiced	voc	frt	semihi	lax
iy	voiced	voc	frt	hi	ten
jh	voiced	frc	palv	nil	nil
k	unvoiced	stp	vel	nil	nil
kcl	unvoiced	stp	vel	nil	nil
l	voiced	latapp	alv	nil	ten
m	voiced	nas	lab	nil	nil
n	voiced	nas	alv	nil	nil
ng	voiced	nas	vel	nil	nil
nx	voiced	flap	alv	nil	nil
ow	voiced	voc	bak	nil	ten
oy	voiced	voc	nil	nil	ten
p	unvoiced	stp	lab	nil	nil
pau	sil	sil	sil	sil	sil
pcl	unvoiced	stp	lab	nil	nil
q	unvoiced	stp	glo	nil	nil
r	voiced	app	alv	nil	ten
s	unvoiced	frc	alv	nil	nil
sh	unvoiced	frc	palv	nil	nil
t	unvoiced	stp	alv	nil	nil
tcl	unvoiced	stp	cor	nil	nil
th	unvoiced	frc	den	nil	nil
uh	voiced	voc	cen	semihi	lax
uw	voiced	voc	bak	hi	ten
ux	voiced	voc	cen	hi	ten
v	voiced	frc	lab	nil	nil
w	voiced	app	lab	nil	ten
y	voiced	app	pal	nil	ten
z	voiced	frc	alv	nil	nil
zh	voiced	frc	palv	nil	nil

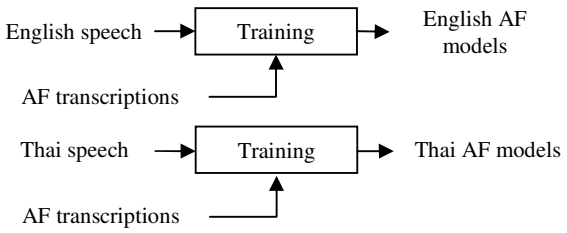

Figure 2: Training AF models.

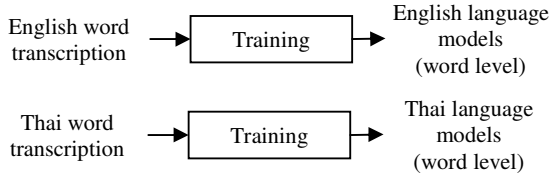

Figure 3: Training language model for each language.

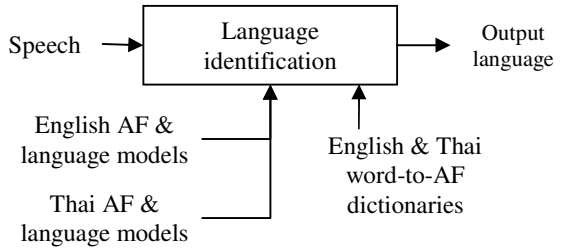

Figure 4: Language identification.

and log energies below a silence floor (set to -50 dB) clamped to that floor.

These 13 front-end parameters are expanded to 39 front-end parameters by appending first and second order differences of the static coefficients. The chosen parameters have been used extensively [6] and have proved to be one of the best choices for HMM-based speech recognition systems.

In this paper, two transcription types are used: phoneme transcriptions and AF transcriptions. Phoneme transcription is used to train phoneme models as a baseline while AF transcription is used to train AF models. *Temporal* training technique is then used for model initialization according to the transcription. When time-aligned transcriptions are provided, each model can be trained separately according to the time boundary given in the transcriptions. The advantage of this technique is that more explicit models can be initialised which can lead to higher model accuracy. However, a lot of highly accurate transcriptions are also required for this technique. Each model contains 5 states and the covariance matrices of all states are diagonal. Fig. 1 shows a 5-state left-right HMM as used in the systems.

Maximum likelihood estimators are used to train HMM parameters [7]. The number of training iterations after each change is determined automatically in line with [8]. After context-independent AF models are trained, they are expanded to context-dependent models using a backing-off technique [9]. The context-independent phoneme models are expanded to context-dependent models using a tree-based state tying technique [10]. Only 1-mixture models are used in order to reduce system complexity. Fig. 2 illustrates AF model training diagram.

The language models are trained from the training set on each language (word level) using a back-off bigram. For the recognition process, the Viterbi algorithm is used without any pruning factor. The recognizer for each language is run in parallel and the output with the highest likelihood score is selected. Fig. 3 illustrates language model training diagram for each language while Fig. 4 illustrates the language identification process.

Table 2: Experimental Results

tier	English (%)	Thai (%)	# states
phone	100	100	1,849+5,718
voice	100	97.3	84+108
manner	100	98.2	855+768
place	99.5	92.9	1,959+489
height	100	99.1	333+111
type	100	99.1	258+132

3. Experiment

Kirchhoff, et al. have proposed AF-based language identification systems [11] which differ from the systems in this paper. Their systems identify a language by using information from every *tier* while this paper is trying to use only information on a single *tier* in order to reduce system complexity. As a result, the aim of this experiment is to find the most appropriate *tier* for language identification. The experiment starts by using normal speech recognition (phoneme level) for language identification. Then, AF models on each *tier* are tested. The language identification systems were tested with the two corpora described in the previous section.

Table 2 shows the experimental results. The table shows the language identification accuracy for each *tier* and each language. English (%) column is the identification accuracy of English as English. Thai (%) column is the identification accuracy of Thai as Thai. *Phone tier* is the language identification system with phoneme models. This gives 100% for both English and Thai language accuracy. For the *voice tier*, the English language accuracy is still 100% while Thai language accuracy drops to 97.3%. This means in 100 Thai sentences, 2.7 sentences are identified as English. The language identification from *manner* models is better than *voice* models. The accuracies for English are equal while the *manner* model accuracy rises to 98.2% for Thai. *Place* models are not as good for language identification. The English accuracy is only 99.5% and the Thai accuracy is only 92.9%. *Height* and *Type* models gain similar accuracies (100% for English and 99.1% for Thai). The *#states* column shows the number of states on each *tier*. The numbers before plus sign are the number of states for English context-dependent models. The numbers after plus sign are the number of states for Thai context-dependent models.

From Table 2, the highest accuracy models for language identification are phoneme models. This is not surprising

since phoneme models contain more higher-level information than AF models. However, the number of states for phoneme models is 7,567 in total (1,849 for English and 5,718 for Thai). This is computationally expensive. The lowest number of states is found on *voice tier* with only 84 states for English and 108 states for Thai. The number of states on the *manner tier* is 1,623 (855 for English and 768 for Thai) which is considerably higher than the number of states on *voice tier* yet the accuracy of the system as a whole is better. While these three results indicate that higher the number of states results in higher language identification accuracy, this is not always true; the accuracy and the number of states on *place tier* demonstrate the opposite result. The number of states for *place* models (2,448) is higher than either the number of states for *manner* or *voice* models. Conversely, the accuracy of *place* models is lower than the accuracies from these two types of models. Therefore, the number of states has no obvious impact on the language identification accuracy.

The *height* and *type* models are quite good for the language identification task. The numbers of models are relatively low (444 on *height tier* and 390 on *type tier*) and the accuracies are relatively high. On *height* and *type* tier, all consonants are converted to *nil* features. This makes these *tiers* contain less consonant information. The higher accuracy on less-consonant-information *tier* means consonants provide less language information than vowels.

4. Conclusions

This paper presents an alternative approach to HMM-based language identification whereby AF models are used instead of phoneme models. The benefits of the AFHMM-based language identification approach can be summarized as follows. Firstly, since AFs are typically similar in many languages, portability from one language to another is easier. Secondly, the smaller number of AFs makes the processing time for AFHMM-based systems shorter than the time required for phoneme-based systems.

From the experimental results, the most appropriate AF models are on *type tier* (100% accuracy for English, 99.1% accuracy for Thai, only 390 states in the system).

The implementation of AFHMM-based language identification is very easy since they are almost the same as HMM-based speech recognition systems. In AFHMM-based language identification systems, HMMs are trained from AF transcriptions instead of from phoneme transcriptions. The phoneme lexicon is also converted into an AF lexicon according to a feature table. The models for each language are trained independently. During recognition, two recognizers are run in parallel and the winning output identifies the language in question.

There are two specific plans for future work. Firstly, since the *height* and *type* models demonstrated good results in Thai and English language identifications, the models on these *tiers* will be tested with more languages in order to show that the result from these experiments are not just only for Thai-English language identification. Secondly, AF-based systems will be re-implemented in more complicated fashion. All phonemes will not be converted according to only on one *tier*. Some algorithm will be implemented to determine what feature is most suited to a particular phoneme and language. For example, *low* might contain more language-specific information for *aa* while *frt* might contain more language-specific information for *ae*.

5. Acknowledgements

This material is based upon works supported by the Science Foundation Ireland for the support under Grant No. 02/IN1/I100. The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

6. References

[1] Zissman, M. A. and Berkling, K. M., "Automatic Lanugage Identification", *Speech Communication, Vol. 35, 2001, p 115-124.*

[2] Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., DARPA *TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM,* NIST, 1993.

[3] Kasuriya S., Sornlertlamvanich V., Cotsomrong P., Jitsuhiro T., Kikui G. and Sagisaka Y., "NECTEC-ATR Thai Speech Corpus", *In Proc. International Committee for Co-ordination and Standardisation of Speech Databases*, 2003.

[4] Kanokphara S. and Carson-Berndsen J., "Automatic Question Generation for HMM State Tying using a Feature Table", *In Proc. The Tenth Australian International Conference on Speech Science & Technology*, 2003.

[5] HTK, *http://htk.eng.cam.ac.uk/*

[6] Davis, S.B., Mermelstein, P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Trans. Acoustic Speech and Signal Processing, Vol. 28(4), 1980, p 357-366.*

[7] Juang, B.H., "Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains", *AT&T Tech. J., Vol. 64(6), 1985.*

[8] Tarsaku, P., Kanokphara, S., "A Study of HMM-Based Automatic Segmentations for Thai Continuous Speech Recognition System", *In Proc. Symposium on Natural Language Processing*, p 217-220, Thailand, 2002.

[9] Kanokphara, S. and Carson-Berndsen, J., "Better HMM-Based Articulatory Feature Extraction with Context-Dependent Model", *In Proc. The 18th International Florida Artificial Intelligence Research Society Conference*, 2004.

[10] Odell, J., "The Use of Context in Large Vocabulary Speech Recognition", *Ph.D. Thesis*, Cambridge University, Cambridge, 1995

[11] Kirchhoff, K. and Parandekar, S, "Multi-Stream Statistical Language Modeling with Application to Automatic Language Identification", *In Proc. Eurospeech*, 2001.