MULTI-LINGUAL PHONEME RECOGNITION EXPLOITING ACOUSTIC-PHONETIC SIMILARITIES OF SOUNDS

Joachim Köhler

Siemens AG, Munich, Germany
E–mail: Joachim.Koehler@zfe.siemens.de

ABSTRACT

The aim of this work is to exploit the acoustic-phonetic similarities between several languages. In recent work cross–language HMM-based phoneme models have been used only for bootstrapping the language–dependent models and the multi–lingual approach has been investigated only on very small speech corpora. In this paper, we introduce a statistical distance measure to determine the similarities of sounds. Further, we present a new technique to model multi-lingual phonemes. The experiments are conducted with the OGI Multi-Language Telephone Speech Corpus for the languages American English, German and Spanish. In the first experiment phoneme recognition rates between 39.0% and 53.9% are achieved using language–dependent models. Using cross–language models yields for some phonemes improvement, but in average a degradation of recognition performance is observed. However, cross–language models speeds up the cross–language transfer and reduces the size of the phoneme inventory of multi-lingual speech recognition systems. Finally, a new method of modelling multi-lingual phonemes, which can be used for a variety of language, is presented. This technique reduces the number of phoneme-based units in a multi-lingual speech recognition system.

1. INTRODUCTION

The multi–lingual approach addresses two aspects of a multi–lingual speech recognition system. First, there is a demand to optimize the process of cross–language transfer of speech recognition technology. Porting a speech recognition system from one language to another is an expensive and time consuming process. At the moment, state of the art cross–language transfer requires collection of a huge database in the new language and a complete new training of the acoustic models. For a robust and commercial telephone–based application utterances of at least 1000 speakers have to be collected and evaluated. Further the training of the acoustic Hidden Markov Models (HMMs) takes a lot of time and requires a lot of computational power. To reduce cost and time of cross–language transfers multi–lingual phoneme modelling becomes an important issue for international companies and speech laboratories.

Second, many applications are limited in memory and computational resources. For a multi–lingual speech recognition system which covers a variety of languages it can be necessary to reduce the number of parameters to fulfil the hardware limitations. Therefore, acoustically similar models and parameters should be used for different languages.

In our multi–lingual approach the similarities of sounds across languages are exploited. The fact that there are similarities of sounds is also documented in international phonetic inventories, like IPA, SAMPA or Worldbet. These inventories are developed by phoneticians classifying the sounds of many languages. This classification was done by sophisticated phonetic knowledge rather than by statistical measurement. In this work we present a statistical method to determine the similarities of sounds across languages.

2. PREVIOUS WORK

The idea of creating multi–lingual phoneme models was first presented in [3]. In this paper the terms poly– and mono–phonemes were introduced. Poly–phonemes are sounds whose realisational properties across several languages are similar enough to be equated. Mono–phonemes are sounds which have language–dependent properties. The experiments were conducted with the EUROM 0 database containing 8 minutes of speech for each language[9]. In another work the cross–language transfer of a speech recognition system from English to Japanese was speeded up by using sounds of the source language for bootstrapping and for adaptation of the HMMs of the target language[2]. Several multi–lingual recognition tasks were presented using language–dependent models. The recognition performance for several languages on phonetic and word level were compared in [6] and [8].

3. MULTI–LINGUAL PHONEME MODELLING

3.1. OGI–ML Speech Corpus

For our investigations we used the OGI Multi-Language Telephone Speech Corpus [10]. The languages American English, German and Spanish were selected from the corpus which covers 11 languages in all. For each language at least 100 speakers were collected. The utterances (story-before-tone) contain about 60 seconds of unconstrained spontaneous speech and were spoken once by each speaker. The sentences

Language	#speak.	#train	#test	#nar_ph	#br_ph
English	149	100	49	66	41
German	100	75	25	72	44
Spanish	102	76	26	60	40
Total	351	251	100	198	125

Table 1: number of speakers (= utterances) for training and test of the OGI-MLTS corpus. Further, number of phonetic units for each language (narrow and mapped broad transcription)

are labelled with the Worldbet phonetic inventory [5]. To reduce the number of phonetic units the diacritic part of the labels were removed. Further, the labels were mapped to a reduced phoneme set to guarantee a reliable estimation of the parameters. These two steps yield a reduction from 198 narrow based phonetic units(nar_ph) to 125 broad phonemes (br_ph).

Alternative to the OGI MLTS–Corpus there are many language–dependent speech corpora, like PHONDAT(German), TIMIT and Wall Street Journal(English) or Bref(French). These corpora have the advantage that they contain much more speech material than the OGI corpus and that it is possible to train context–dependent phoneme models. The disadvantage using these huge corpora is, that all these databases were recorded with different equipment and that they differ in speaking style and quality. So the comparison of phoneme models across languages can be overlapped and covered up by different recording conditions.

3.2. International Phonological Inventories

The most used phonetic inventory is the International Phonetic Alphabet (IPA). Because IPA contains non ASCII symbols which are difficult to process on different computer platforms, some alternative phonetic inventories were created. For European languages SAMPA was invented and designed in the scope of projects in the European Union (ESPRIT). SAMPA contains a broad phoneme set which uses the same symbol for quite different sounds across languages (i.e r-sounds). The Worldbet ([5]) in its basic form is an ASCII representation of the IPA. Further, it is possible to provide each symbol with diacritics. These diacritics allow to mark allophonic realizations of the phonemes. At the current status there are 299 different phonetic symbols for a variety of different languages.

3.3. Language–Specific Phoneme Properties

In the beginning we defined the goal to exploit and model similar sounds. However, the realization of the phonemes in each language can differ. The reasons for the different acoustic realization are:

- different phonetic context (because of different phoneme sets)
- different speaking styles
- different prosodic features
- different allophonic variations

One important aspect is called *principle of sufficient perceptual separation*[7]. This means that the sounds of a language are kept acoustically distinct so as to make it easier for the listener to distinguish one from another. Because each language has a separate phoneme inventory, the boundaries between two similar phonemes in each language are language–specific. Hence, the variation of the realization of a sound has a language–specific component.

3.4. Statistical Phoneme Modelling

The phonemes are modelled by continuous density Hidden Markov Models (CD-HMM) [4]. As density functions Laplacian mixtures are used. Each phoneme consists of a 3 state left-to-right HMM. The acoustic feature vectors consists of 24 mel-scaled cepstral, 12 delta cepstral, 12 delta delta cepstral, energy, delta energy and delta delta energy coefficients. The length of the analysis window is 25 msec and the displacement is 10 msec for each frame. Because of the limited size of the speech corpus only context independent phoneme models are created.

3.5. Distance Measure

One important issue of this investigation is to find a reliable distance measure for phonemes modelled by a 3 state Markov model. This distance measure can be used for clustering or substitution of multi–lingual phoneme models. It is also useful for developing or evaluating a multi–lingual phonetic inventory, like IPA, SAMPA or Worldbet.

To measure the distance or the similarity of two phoneme models we use a relative entropy-based distance metric [1]. During training the parameters of the mixture Laplacian density phoneme models are estimated. Further, for each phoneme a set of phoneme tokens X is extracted from a test or development corpus. A phoneme token is a realization or observation of a phoneme and is marked by the phonetic label.

Given two phoneme models λ_i and λ_j and given the sets of phoneme tokens or observations X_i and X_j the distance between model λ_i and λ_j is defined by:

$$d(\lambda_i, \lambda_j) = \log p(X_i|\lambda_i) - \log p(X_i|\lambda_j)$$

This distance measure can be considered as log likelihood measure which tests how well two different models fit to the same data X_i. Corresponding, the distance between model λ_j and λ_i is defined by:

$$d(\lambda_j, \lambda_i) = \log p(X_j|\lambda_j) - \log p(X_j|\lambda_i)$$

To get a symmetric distance the average is taken:

$$d(\lambda_i; \lambda_j) = \frac{1}{2}(d(\lambda_i, \lambda_j) + d(\lambda_j, \lambda_i))$$

4. PHONEME RECOGNITION

The phoneme models were trained by a standard viterbi–based maximum likelihood training algorithm. The recognition tests were performed using the phonetic labels resulting

Language	#Tokens	LDP[%]	ML1[%]	ML2[%]
English	21191	39.0	37.3	37.0
German	9430	40.0	34.7	37.7
Spanish	9525	53.9	46.0	51.6
Total	40146	42.8	38.8	40.8

Table 2: phoneme recognition rates depending on number of phonemes and densities; #Tokens: number of classified tokens; LDP: 125 language dependent models using 9696 densities; ML1: 72 multi-lingual models using 6419 densities; ML2: 72 multi-lingual models using 6155 densities

in an isolated rather than a continuous phoneme recognition task.

4.1. Language–Dependent Models

In the first experiment we conducted training and test for each language separately to get baseline recognition results. Hence, the phoneme models are language-dependent. The phoneme recognition results are listed in Table 2 (column: LDP).

It is obvious that the phoneme recognition rate for Spanish is much higher than for German and English. This can be explained by the simple vowel structure of Spanish. In German and English short and long vowels exist which are difficult to distinguish during labelling as well as recognition.

The recognition rate increased using automatically aligned labels generated by the forced viterbi algorithm instead of using the original hand labelled transcription. Using the re–aligned labels for recognition the accuracy increased to 49.2%, 49.9% and 62.0% for English, German and Spanish, respectively.

4.2. Cross–Language Phoneme Model Substitution

In the second experiment we evaluated the performance of the English phoneme models used instead of the German phoneme models in a German recogniser. Therefore, we computed the distance $d(\lambda_{GE}, \lambda_{AE})$ between the German (GE) and English (AE) phoneme models using the proposed distance metric. Then we substituted the German phonemes with the corresponding English phonemes which had the smallest distance or highest similarity to each other. The phoneme recognition results are presented in column 4 of Table 3. For a few phonemes this substitutions yield improvement (/k/, /p/ and /N/). This is also shown by the negative distance values $d(\lambda_{GE}, \lambda_{AE})$ between German and English models. A negative distance value means that the English model generates a higher likelihood than the German model. However, most of the substitution yields a significant degradation. In spite of the fact that the English models were trained with more data than the German models, the language–specific models achieve in average a higher performance than the substituted cross–language models.
The ability to use this distance measure to evaluate a phonetic inventory can be demonstrated for the diphones /aU/ and /aI/. A high dissimilarity between the German and the English model for /aU/ was observed meaning that both sounds should get a different symbol in a multi–lingual inventory. On the other hand the German and English diphthong /aI/ showed a high similarity justifying the use of the same symbol for these two sounds.

Phoneme	$d(\lambda_{GE}, \lambda_{AE})$	GE[%]	AE[%]	Multi-Ph.[%]
n	0.995	60.9	48.6	56.4
m	-0.625	44.7	40.4	49.3
N	-2.041	22.2	26.4	23.6
p	-1.937	25.0	30.6	25.0
b	2.731	34.5	29.1	32.4
d	0.484	27.7	17.0	23.4
t	0.488	45.7	35.5	36.6
g	4.207	30.2	15.3	32.0
k	-5.368	37.9	53.8	38.7
s	2.427	45.4	34.1	29.5
z	5.981	40.4	21.3	42.6
f	1.344	52.2	46.5	45.2
h	3.315	50.0	20.1	25.0
S	6.654	54.9	38.2	52.9
aI	1.722	63.0	56.3	56.7
aU	15.35	53.4	5.5	42.5

Table 3: phoneme recognition rates for German consonants and diphthongs; col. 2: distance between German and English phoneme models; col. 3,4,5: recognition rate using German, English or Multi–Lingual models, respectively. The phonetic symbols are taken from the Worldbet

4.3. Multi–Lingual Phoneme Models

The main studies have focussed on the ability to create multi–lingual phoneme models which can be used in a variety of languages. For each symbol of a multi–lingual phonetic inventory a separate statistical model should be created. However, the previous experiments have shown that the language–dependent models yield a higher performance than the cross–language models. Now we want to combine the language–dependent and the language–specific acoustic properties to a multi–lingual model. In [9] the poly–phonemes are defined as phonemes which are similar enough to be modelled them as one phoneme. A drawback of this approach is that the complete acoustic space of a poly–phoneme model is used during a language–dependent recognition test. In our approach we identify and model regions of the acoustic space in which similar phonemes are overlapping. Therefore, we apply an agglomerative density clustering technique to reduce equal or similar realizations of similar phonemes. Only the densities of corresponding states in the phoneme are clustered.

Figure 1 shows the architecture of a single multi-lingual phoneme. Densities used in all languages are located in the hatched region. Whereas the densities are tied across different languages, the mixture weights of the densities are language–dependent. This property should emphasize the fact, that a specific realization of a phoneme appears in one language more often than in another language. The density clustering was conducted with different cluster thresholds.

Thr.	#densit(a,b,c).	Engl.[%]	Germ.[%]	Span.[%]
0	341(0 0 341)	46.7	44.7	59.4
2	334(0 14 327)	45.0	46.4	57.5
3	303(27 34 280)	48.0	45.8	57.5
4	227(106 57 178)	50.9	44.1	58.7
5	116(221, 48, 72)	49.3	43.1	57.0
6	61(285, 22, 34)	41.2	38.6	50.4

Table 4: recognition rates for phoneme /m/ using different cluster thresholds. (a,b,c) denotes the number of densities clustered to the 3, 2 or 1 language region, respectively. Method ML2 is used

Number of densities and recognition rate for the phoneme /m/ are given in Table 4. Using a cluster threshold of 5 the number of densities were reduced by a factor of 3 without any significant degradation. In this case 221, 48 and 72 of the 341 initial densities are clustered to the poly–phoneme region, to the two–language and to the mono language region, respectively.

The recognition rates for the complete multi–lingual system are given in column 5 and 6 of Table 2 (ML1, ML2). For the experiment ML1 the conventional poly–phoneme definition is used meaning that the complete acoustic region of a poly–phoneme (outer contour of Figure 1) is used for recognition. The new proposed method including only partial overlap of acoustic region yields improvement of 2.0% (ML2). However, the recognition rates are lower than in the language–dependent case.

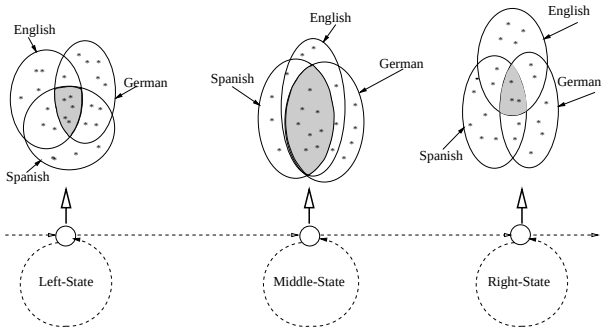

Figure 1: Multi–lingual 3 state left–to–right hidden Markov model. Filled region shows densities (in a 2 dimensional space) used in all of the 3 languages (poly–phoneme region).

5. CONCLUSIONS AND FUTURE WORK

In this paper, we described a method to determine the similarities of sounds across different languages. Further, a new approach of multi–lingual phoneme modelling was introduced. The proposed acoustic–phonetic modelling considers language–dependent as well as language–independent properties using a density clustering algorithm.

However, many questions have to be clarified. First, the modelling and density clustering of the multi–lingual phoneme models have to be improved and optimized. Furthermore, this technique has to be applied to context-dependent phoneme models (diphones or triphones). It is expected, that the overlap of densities and cross–language similarities between context–dependent phoneme models are higher than for context–independent models. However, this requires multi–lingual databases of a huger size than the OGI corpus offer. Hence, we have to evaluate this approach on different databases, like TIMIT and PHONDAT. Channel compensation and normalisation technique may help to reduce the database dependencies. Finally, additional languages have to be incorporated and evaluated for this multilingual approach.

6. ACKNOWLEDGEMENTS

The author would like to thank Dr. Höge for his help and encouragement.

7. REFERENCES

1. V. Digalakis A. Sankar, F. Beaufays.: "Training Data Clustering For Improved Speech Recognition.", In *Proc. EUROSPEECH '95*, pages 503 – 506, Madrid, 1995.
2. W. Anderson, B. Wheatley, K. Kondo and Y. Muthusamy.: "An Evaluation of Cross-Language Adaptation for rapid HMM Development in a new Language.", In *Proc. ICASSP '94*, pages I237 – I240, Adelaide, 1994.
3. P. Dalsgaard and O. Andersen.: "Identification of Mono- and Poly-phonemes using acousitc-phonetic Features derived by a self-organising Neural Network.", In *Proc. ICSLP '92*, pages 547 – 550, Banff, 1992.
4. A. Hauenstein and E. Marschall.: "Methods for Improved Speech Recognition Over the Telephone Lines.", In *Proc. ICASSP '95*, pages 425 – 428, Detroit, 1995.
5. J. L. Hieronymus.: "ASCII Phonetic Symbols for the World's Languages: Worldbet.", preprint, 1993.
6. J. Salavedra, C. Jacobsen, M. Rahim, I. Zeljkovic and J.G. Wilson: "Multi-lingual Connected Digits Recognition.", In *Proc. EUROSPEECH '95*, pages 2119 – 2122, Madrid, 1995.
7. P. Ladefoged: "*A Course in Phonetics.*", *Harcourt Brace Jovanovich*, San Diego, 1993.
8. L.F. Lamel and J.L. Gauvain.: "Cross-lingual Experiments with Phone Recognition.", In *Proc. ICASSP '93*, pages II507 – II501, Minneapolis, 1993.
9. P. Dalsgaard O. Andersen and W. Barry.: "Data-driven Identification of Poly- and Mono-phonemes for four European Languages.", In *Proc. EUROSPEECH '93*, pages 759 – 762, Berlin, 1993.
10. A. Cole Y.K. Muthusamy and B.T. Oshika.: "The OGI Multi-language Telephone Speech Corpus.", In *Proc. ICSLP '92*, pages 895 – 898, Banff, 1992.