LANGUAGE ADAPTATION OF MULTILINGUAL PHONE MODELS FOR VOCABULARY INDEPENDENT SPEECH RECOGNITION TASKS

Joachim Köhler

Siemens AG, Corporate Technology, D-81730 Munich, Germany
Joachim.Koehler@mchp.siemens.de

ABSTRACT

This paper presents our new results on multilingual phone modeling and adaptation into a new target language which is not included in the trained multilingual models. The experiments were carried out with the SpeechDat(M) and MacroPhone databases including the languages French, German, Italian, Portuguese, Spanish and American English. First, we constructed language-dependent and multilingual phone models. The recognition rate for an isolated word task decreased in average only by 3.2% using 95 multilingual instead of 232 language-dependent models. Second, we investigated adaptation techniques for cross-language transfer and showed that only 100 utterances from a new language were needed for adaptation. Using the MAP algorithm the recognition rate was improved from 79.9% to 84.3%. Finally, we defined a phonetic based dissimilarity measure between 2 languages and compared language-dependent and multilingual models for the purpose of cross-language transfer.

1. INTRODUCTION

Over the last years automatic speech recognition systems have reached a level which allows the introduction of commercial products. However, a new problem has occurred: the language-dependency of current recognition technology. The phonetic models used in state-of-the-art system are extremely language-dependent. The overall goal of our research activities is to create a multilingual and almost language independent recognition system which works in the most important languages of the world. We started our multilingual approach with OGI MLTS database [1] based on the work of [2]. Nowadays, even larger multilingual databases are available like SpeechDat(M)[1], Call-Home etc. These databases allow a robust modeling of phonetic units for different languages.

However, for each new language huge database collections has to be carried out. This is an expensive and time consuming procedure. Alternativly we apply adaptation techniques to reduce the amount of training data for cross language transfer. In previous work we combined the multilingual approach with task adaptation to bootstrap and adapt a digit recognizer for the Slovene language [3]. In this paper we investigate the cross-language transfer of multilingual phone models for vocabulary and database independent tasks. Training and adaptation is carried out with phonetically balanced speech material.

The paper is organized as follows: First we describe the modeling of multilingual phones. The evaluation of the models are carried out on isolated word and phoneme level for 6 languages. In the following section we investigate adaptation and bootstrapping strategies for a new target language. Therefore, we compare different methods depending on the size of the adaptation material. Another focus of our investigations is the influence of the similarities of languages for adaptation. Therefore, we define a dissimilarity measure between languages.

2. MULTILINGUAL SYSTEM FOR TELEPHONE SPEECH

2.1. Multilingual Speech Databases

For research we used the telephone speech databases SpeechDat(M) and MacroPhone. The SpeechDat(M) database covers 8 languages: French (FR), Italian (IT), British-English (BE), Portuguese (PT), German (GE), Spanish (SP), Swiss-French (SF) and Danish (DA). Each language contains utterances of 1000 speakers recorded over the telephone line. The speakers have spoken isolated and connected digits, spelled letters, applications words, phonetically balanced sentences etc. For our investigations we used the 5 languages GE, FR, PT, IT, SP. The 1000 speakers are divided in sets of 666, 167, 166 for training, development and testing, respectively. Each corpus is delivered with a phonetic lexicon which covers all words spoken in the database. The lexicon entries are transcribed with the SAMPA notation. All utterances contain a orthographic transcription on word level. The phonetic labels were generated automatically during the training procedure using the Viterbi-algorithm.

[1] For information about SpeechDat see the following URL's:
http://www.phonetik.uni-muenchen.de/SpeechDat.html
http://www.icp.grenet.fr/ELRA/home.html

	#speakers	#utterances		hour.min		#Phone-
	tr- dev- te	TR-ALL	TR-PHO	TR-ALL	TR-PHO	Units
FR	667- 166- 167	30.6K	6.0K	15.37	5.03	37
GE	667- 166- 167	22.3K	5.0K	12.08	4.18	38
IT	667- 166- 167	26.2K	5.8K	15.27	4.15	49
PT	667- 166- 167	26.4K	5.9K	17.09	7.33	38
SP	667- 166- 167	28.0K	6.0K	16.32	5.38	31
AE	1000- 500- 500	39.7K	6.4K	19.01	5.12	39
All	4335-1330-1335	173.2K	35.1K	95.54	31.59	232

Table 1: Organisation of the databases. (TR-ALL: #utt. of the complete training set; TR-PHO: #utt of the phonetically balanced sentences) (hour.min: pure speech duration in hours and minutes without silence)

Additionally, we used for American English (AE) 1000 speakers of the MacroPhone database which is similar structured than SpeechDat(M). One important difference is that the MacroPhone database is not delivered with a phonetic lexicon. Hence, we extracted the phonetic transcription from the CMU lexicon which has as phone inventory a subset of the TimitBet. Because MacroPhone has a separate test set (devtst) no splitting of the database was necessary. Table 1 summarises the division and organisation of the multilingual databases used for the following experiments.

2.2. Multilingual Phone Modelling

The phones were modelled by continuous Gaussian mixture densities (CD-HMM). Each phone consisted of a 3 segment left-to-right HMM. The segments again had 2 tied states sharing the same mixture density. The acoustic feature vectors consisted of 24 mel-scaled cepstral, 12 delta cepstral, 12 delta delta cepstral, high pass filtered energy, delta energy and delta delta energy coefficients. The length of the analysis window was 25 msec and the displacement was 10 msec for each frame. Further, the feature vectors were transformed by a LDA [6]. Using LDA the number of coefficients of the feature vector were reduced from 51 to 24. The LDA was trained on the basis of context-independent multilingual phone models. The classes which should be discriminated were the segments of the multilingual phone set. Hence, we worked with 1 multilingual LDA rather than having a LDA for each language. Previous experiments have shown that this simplification has no significant influence on the recognition performance.

In the first step we trained monolingual models for all 6 languages. Each language has some properties which has to be taken into account. For example French contains elisions and liasons. American-English is transcribed with a subset of the TimitBet instead of SAMPA. Also for German we used a reduced set of SAMPA which is called SPICOS.

The first set of models were trained with all utterances defined in column 5 of table 1 (TR-ALL). This means that also the isolated and specialized words were included in the training set. Hence, there was an overlap between the vocabulary of the training and test set. To achieve real vocabulary independent models only the phonetically balanced sentences defined in column 6 of table 1 (TR-PHO) were used to train the second set of language-dependent models. This yielded in a reduction by the factor of 3 of the training material. For both sets we ended up with 232 monolingual phone models plus 3 models for non speech events namely, pure silence [si], any kind of background noise [nib] and unknown speech [unk]. The number of densites of the CDHMM for each language varied between 4K and 6K depending on the number of phones in each language. In total 31K densities were used for the 6 languages.

Finally, we mapped all language-dependent models to their corresponding IPA symbol [5]. Hence, we reduced the number of 232 monolingual to 95 multilingual phones. This mapping yielded in a HMM containing 13K densities. The training was performed with the 35.1K utterances (almost 32 hours pure speech) of the phonetically balanced part of the training set. All models were context-independent and were trained with the Viterbi based Maximum Likelihood training procedure. The number of densities

2.3. Recognition Tests for the 6 languages

The evaluations of the 6 languages were carried out on isolated word and on phonetic level. The application word task (APPL-task) contains all words defined in the in the SpeechDat(M) database (”a”-sentences) for general telecommunications applications like ”operator”, ”information”, ”record”, etc. The vocabulary size of the APPL-task varied from 47 (IT) to 70 (SP) for SpeechDat(M). For AE there were no core application word vocabulary but a huge list of command words (685 entries) which increased the recognition perplexity. For phonetic decoding (PHONE-task) we used a simple bigram phone-based language model which were trained with the training set (TR-PHO). The given phone accuracy considers substitutions, deletions and insertions. The tests were performed for the monolingual models trained with TR-ALL (LDP-ALL), for monolingual models trained with TR-PHO (LDP-PHO) and for the multilingual models trained with TR-PHO (IPA-PHO).

The results presented in table 2 needs some interpretation. The performance of the APPL-task in the languages GE, IT and SP benefits from including the test vocabulary in the training set. But in general this has no significant impact to the recognition results (in average 89.3% vs. 89.0% for APPL-task). For phonetic decoding the isolated words even hurt the system (46.8% vs. 48.6%). German shows the best performance for the APPL-task. This is probably due to the highly optimized lexicon we had for our native language German. Further, it is obvious that the phone recognition rate for American-English is the lowest. We explain this effect that AE is spoken less accentuated than Italian or Spanish. Also the more complicated grapheme to phone mapping in AE may be a reason. In average the multilingual approach yields in a degradation of 3.2% (APPL) and 4.9%

Task	Lang.	#Rec-. Tokens	Lex- Size	LDP- ALL	LDP- PHO	IPA- PHO
APPL	FR	1420	57	91.3%	92.2%	90.9%
	GE	949	49	97.6%	96.6%	91.6%
	IT	983	47	95.1%	94.4%	93.6%
	PT	931	61	93.2%	93.0%	89.6%
	SP	1242	70	94.3%	93.3%	92.5%
	AE	2612	685	64.5%	64.9%	56.5%
	av	–	–	89.3%	89.0%	85.8%
PHONE	FR	12964	37	42.2%	48.3%	42.2%
	GE	12839	38	46.1%	48.5%	41.9%
	IT	10804	49	51.1%	53.2%	47.9%
	PT	21751	38	46.7%	47.0%	42.2%
	SP	17512	31	56.6%	56.9%	54.7%
	AE	10815	39	38.0%	37.7%	33.3%
	av	–	–	46.8%	48.6%	43.7%

Table 2: Results with language-dependent and multilingual models for 6 languages

(PHONE) compared to the monolingual models. However, the number of context-independent phone models was reduced from 232 to 95 and the number of density functions from 31K to 13K. Hence, it is possible to build a multilingual system with language-independent models, especially when the number of system parameters is limited.

3. CROSS-LANGUAGE ADAPTATION

3.1. Scenario for cross-language transfer

In the previous section we demonstrated the usefulness and feasibility of multilingual phone modeling. In this section we exploit the multilingual models for cross-language transfer. Therefore, we excluded our native language German from the multilingual model set which contained for the next series of experiments the speech material of the 5 languages AE, IT, FR, PT and SP. The main reason for the selection of German was that we wanted to test the cross-language models with a another test database which means different channel characteristics and different vocabulary. Both effects yields in lower recognition performance. However, we try to build HMMs which perform robustly for different environments. Therefore, we tested the cross-language models with the German Voicemail database (VM-62). It consists of 6935 utterances spoken from 140 speakers. The vocabulary size is 62.

3.2. Bootstrapping Methods for the New Target Language

Here, we investigated 3 different methods to build a speech recognition system in a new target language. The 3 methods were:

New-Training (SCRATCH)

The method SCRATCH means that the training was started from the phonetic labels. Hence, the first step of the training was the initialization of the models. After the initialization process 6 iterations of Viterbi-training were performed. For this method the phonetic label files were required. As simplification we assumed that the labels were already existing.

Bootstrapping (BOOT)

The BOOT-method means that the initialization phase could be skipped and that only the phone sequence has to be known. The multilingual seed models served only for providing a good segmentation on state level. For the BOOT-method we ran 2 Viterbi iterations. This usefulness of the BOOT-method was also demonstrated in [7].

MAP-Adaptation (MAP)

There are several methods and applications for adapting models for a new and specialized acoustic environment. Typical applications are speaker and task adaptation. Well known techniques are transformation-based approaches like MLRR and Bayesian adaptation. In our work we concentrated on the MAP adaptation which is applied only for the means μ of the continuous density HMM (CDHMM) [8] [9]. The adaptation is given by:

$$\mu_{s,m} = \frac{\tau \hat{\mu}_{s,m} + \sum_{t=1}^{N_{s,m}} x_t}{\tau + N_{s,m}} \qquad (1)$$

where $\hat{\mu}_{s,m}$ is the density of the multilingual HMM of state s and mixture component m. Equation 1 implies the simplification that each feature vector x_t is emitted from the best fitting mixture component rather from a weighted sum of the mixture components. $N_{s,m}$ is the number of frames assigned to the mixture component m of state s. This formula combines the multilingual mean vectors $\hat{\mu}_{s,m}$ (i.e. multilingual seed model) with the parameters estimated during the Viterbi-training. The value for τ was set to 5. Figure 1 shows the recognition result for the 3 different methods and varying sizes of training utterances achieved on the VM-

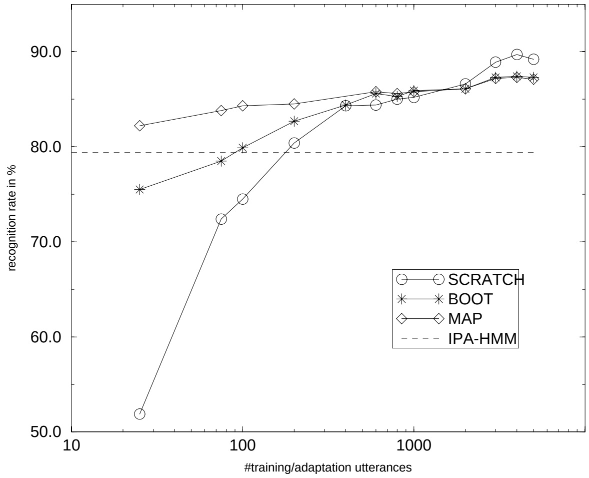

Figure 1: Cross-language transfer with multilingual models

62 database. We performed the tests for 25, 75, 100, 200, 400, 600, 800, 1000, 2000, 3000, 4000 and 5000 utterances, which corresponds to an absolute duration of pure speech without silence of 1, 4, 5, 10, 21, 31, 42, 52, 104, 155, 206 and 258 minutes, respectively. Taking the IPA-based models directly for recognition we achieved 79.4% word accuracy (IPA-HMM). It is obvious that for few utterances the MAP method outperforms the BOOT and SCRATCH method. After 600 utterances (31 min.) all three methods worked equally well. With more than 1000 utterances the SCRATCH method shows the best recognition results. This experiment demonstrated that adaptation is useful for a limited amount of training material. The MAP models combine the robustness of the multilingual models and the details acoustic properties given by the adaptation material.

3.3. Language-dependent versus Multilingual bootstrapping

So far we performed the adaption with multilingual models. In [9], [7] the bootstrapping were carried out with language-dependent models. In this experiment we followed their approach and investigated the influence of the dissimilarity between two languages for bootstrapping and adaption. Therefore we defined a phonetic dissimilarity measure between two languages. This is given by the equations:

$$d(L_1;L_2) = \frac{1}{N_{L1}} \sum_{i=1}^{N_{L1}} d_{min}(\lambda_i;\lambda_{j^*}) \tag{2}$$

$$\lambda_{j^*} = \operatorname*{argmin}_{j=1\ldots N_{L2}} d(\lambda_i;\lambda_j) \tag{3}$$

$$d(\lambda_i;\lambda_j) = \frac{1}{T_{L1}} \sum_{n=1}^{T_{L1}} \log(\lambda_i|X_n^i) - \log(\lambda_j|X_n^i) \tag{4}$$

where N_{L1} and N_{L2} are the numbers of phones in language L1 and L2, respectively. T_{L1} is number of phone tokens X_n^i of model λ_i extracted by using the phonetic labels. $d(\lambda_i;\lambda_j)$ is the distance between the models λ_i and λ_j.

Table 3 shows the dissimilarities of the 5 languages to German. FR and AE are most similar to GE and yields in the best recognition rates in the case of direct transfer of the models (column BASE). Further, we observe a correlation between language dissimilarity and recognition rate. Applying 100 utterances for bootstrapping and adaptation the word accuracy is similar for all languages. Nevertheless, we achieved the best result for all cases with the IPA-based models (trained without GE) which is also emphasized by the low dissimilarity value of 9.7.

4. SUMMARY AND CONCLUSION

In this paper we demonstrated the usefulness and feasibility of the multilingual approach. First, a telephone-based multilingual speech recognition system was built for 6 languages. IPA-based phone modeling reduced significantly the number of parameters in the multilingual environment. Further, we applied the multilingual HMMs for cross language transfer. In combination with MAP adaptation technique it was possible to develop models with a small amount of training data. Using this approach we achieved 84.3% word accuracy on a cross-language and cross-database task with only 5 minutes of pure speech material. Finally, we showed the superiority of the multilingual models against language-dependent models for the purpose of cross language transfer.

LangPair	Dist	BASE	BOOT100	MAP100
FR → GE	21.3	72.0 %	78.2 %	80.9 %
IT → GE	24.7	64.4 %	76.7 %	79.3 %
PT → GE	22.1	67.4 %	77.3 %	78.5 %
SP → GE	23.9	65.5 %	77.4 %	79.3 %
AE → GE	21.7	70.1 %	79.2 %	81.3 %
IPA → GE	9.7	79.4 %	79.9 %	84.3 %

Table 3: Language Transfer with LDP-models and IPA-based models

5. REFERENCES

[1] J. Koehler: “Multi-Lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds”, *Proc. ICSLP’96*, pages 2195 - 2198, Philadelphia, 1996.

[2] 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.

[3] U. Bub, J. Koehler: “In-Service Adaptation of Multilingual Hidden-Markov-Models“ *Proc. ICASSP ’97*, pages 1451 – 1454, Munich, 1997.

[4] P. Bonaventura, F. Gallocchino and G. Micca: “Multilingual Speech Recognition For Flexible Vocabularies”, In *Proc. EUROSPEECH ’97*, pages 355 – 358, Rhodes, 1997.

[5] International Phonetic Association: “The International Phonetic Association (revised to 1993) - IPA chart”, In *Journal of the International Phonetic Association*, 23, vol. 1, 1993.

[6] A. Hauenstein and E. Marschall.: “Methods for Improved Speech Recognition Over the Telephone Lines.“, In *Proc. ICASSP ’95*, pages 425 – 428, Detroit, 1995.

[7] T. Schultz: “Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets”, In *Proc. EUROSPEECH ’97*, pages 371 – 374, Rhodes, 1997.

[8] J.T. Chien, C.H. Lee and H.C. Wang: “Improved Bayesian Learning of Hidden Markov Models For Speaker Adaptation”, *Proc. ICASSP ’97*, pages 1027 – 1030, Munich, 1997.

[9] B. Wheatley, K. Kondo, W. Anderson, Y. Muthusamy: “An Evaluation of Cross Language Adaptation for Rapid HMM Development in a New Language”, *Proc. ICASSP ’94*, pages 237 – 240, Adelaide, 1994.

[10] C. Corredor-Ardoy, J.L. Gauvain, M. Adda-Decker, L. Lamel: “Language Identification With Language-Independent Acoustic Models”, In *Proc. EUROSPEECH ’97*, pages 55 – 58, Rhodes, 1997.