MANUAL AND SEMI-AUTOMATIC APPROACHES TO BUILDING A MULTILINGUAL PHONEME SET

Ekaterina Egorova, Karel Veselý, Martin Karafiát, Miloš Janda and Jan Černocký

Brno University of Technology, BUT Speech@FIT and IT4I Centre of Excellence
xegoro00@stud.fit.vutbr.cz

ABSTRACT

The paper addresses manual and semi-automatic approaches to building a multilingual phoneme set for automatic speech recognition. The first approach involves mapping and reduction of the phoneme set based on IPA and expert knowledge, the later one involves phoneme confusion matrix generated by a neural network. The comparison is done for 8 languages selected from GlobalPhone on three scenarios: 1) multilingual system with abundant data for all the languages, 2) multilingual systems excluding target language 3) multilingual systems with small amount of data for target languages. For 3), the multilingual system brought improvement for languages close enough to the others in the set.

Index Terms— multilingual speech recognition, phoneme set mapping, phoneme confusion matrix

1. INTRODUCTION

The increasing interest in speech-to-speech translation and automatic processing of low-resource languages led to research of multilingual approaches which would ease the system development for a new language. The biggest cost factor in such development is the need of training data for the acoustic model. Several techniques have been investigated to alleviate this problem. The *cross-language transfer* applies a system developed on one language to another one. It has been shown that the performance in new language is proportional to the similarity of the languages [1]. The *language adaptation* technique adapts the system to a new language with only limited data. The performance of the adapted system depends on the amount of available data [2]. When the amount of data becomes sufficient for full training, the *bootstrapping* technique can be used for initializing the new language system by the original one [3].

But having low-cost monolingual systems might not solve the problem completely. To process a recording in an unknown language, it would be necessary to perform language identification on the given recording and then to load the appropriate ASR system. A multilingual system combining the phonetic inventory of several languages into one acoustic model will benefit from total parameter reduction and leaving out the language identification system. Moreover, multilingual systems can switch the languages within one utterance. Further research has shown that such multilingual acoustic models also improve all techniques mentioned above [4, 5, 6]. Additionally, these systems can also better handle foreign accented speech [7].

A multilingual system combining the phonetic inventory of several languages into one acoustic model demands a lot of resources for training because of the size of the joined phoneme set. Moreover, without any preprocessing of the phonetic systems, it may happen that many phonemes appear only in one language and do not add to the multilingual acoustic model. The aim of this work is to show how manual and semi-automatic phonetic approaches to creating a multilingual phoneme set can solve these problems in the multilingual system and help in speech recognition for languages with no or little training data.

2. EXPERIMENTAL SETUP

2.1. Data

The data comes from multilingual database GlobalPhone [8]. The database covers 19 languages with an average of 20 hours of speech from about 100 native speakers per language. This database aims for an acceptable out of vocabulary (OOV) rate in test sets but with occurrences of words from other languages. This requirement was satisfied by newspaper articles which were read by native speakers. The database covers speakers of both genders in ages from 18 to 81 years. The speech was recorded in office-like environment by high quality equipment. We converted the recordings to 8KHz, 16 bit, mono format, to enable usage of telephone data, as in [9].

The following languages were selected for the experiments: Czech (CZ), German (GE), Portuguese (PO), Russian (RU), Spanish (SP), Turkish (TU) and Vietnamese (VI). These languages were accompanied with English (EN) taken from Wall Street Journal database. See Table 1 for detailed numbers of speakers and data partitioning. Each individual speaker appears only in one set. The partitioning followed the GlobalPhone recommendation. The dictionaries for Vietnamese and Russian were obtained from Lingea[1]. The CMU dictionary was used for English.

The data for language models (LM) was obtained from Internet sources (newspaper articles) using RLAT and SPICE tools[2]. The size of gathered corpus for LM training together with the sources are given in Table 2. Bigram LMs were generated for all languages except Vietnamese, which is a syllable language; a trigram LM was created for it.

This work was partly supported by Czech Ministry of Trade and Commerce project No. FR-TI1/034, Technology Agency of the Czech Republic grant No. TA01011328, and by European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070). M. Karafiat was supported by Grant Agency of the Czech Republic postdoctoral project No. P202/12/P604.

[1] http://www.lingea.com

[2] http://i19pc5.ira.uka.de/rlat-dev/index.php, http://plan.is.cs.cmu.edu/Spice/spice/index.php

Lang.	Speakers	Audio	TRAIN	DEV	TEST
GE	77	18	13.2	1.8	1.3
CZ	102	29	26.8	1.2	1.9
EN	311	16	14.2	1.0	1.0
SP	100	22	13.4	1.2	1.2
PO	102	26	14.7	1.0	1.0
TU	100	17	12.0	1.6	1.4
VI	129	19	14.7	1.2	1.3
RU	115	22	16.9	1.3	1.4

Table 1. Number of speakers and amount of audio material in hours overall, for training, development and testing

Lang	OOV	Dict Size	LM Corpus Size	WWW Server
GE	1.92	375k	19M	www.faz.net
CZ	3.08	323k	7M	www.novinky.cz
EN	2.30	20k	39M	WSJ - LDC2000T43
SP	3.10	135k	18M	www.aldia.cr
PO	0.92	205k	23M	www.linguateca.pt/ cetenfolha
TU	2.60	579k	15M	www.zaman.com.tr
VI	0.02	16k	6M	www.tintuconline.vn
RU	1.44	485k	19M	www.pravda.ru

Table 2. Detailed information about language models and test dictionaries for individual tasks.

2.2. Recognition system

The recognition system is based on HMM cross-word tied-states triphone acoustic models. The models contain ≈3000 tied states with 18 Gaussian mixtures per state. Models for each parameter set were trained from scratch using mixture-up maximum likelihood training.

Mel-filter bank based PLP coefficients were used as features. There were 13 direct parameters augmented with deltas and double-deltas totaling in feature vectors with 39 coefficients. Cepstral mean and variance normalization was applied on speaker basis. The resulting models were used for forced alignment of the data. The results for each language trained separately (in terms of WER) are shown in Table 3, column Baseline.

3. IPA USABILITY

The *International Phonetic Alphabet (IPA)*[3] is an alphabetic system of phonetic notation based primarily on the Latin alphabet. It was devised by the International Phonetic Association as a standardized representation of the sounds of spoken language. The general principle of the IPA is to provide one letter for each distinctive sound (speech segment). This means that it does not use combinations of letters to represent single sounds, the way English does with "sh" and "ng", or single letters to represent multiple sounds the way "x" represents /ks/ or /gz/ in English. There are no letters that have context-dependent sound values, as "c" does in English and other European languages, and finally, the IPA does not usually have separate letters for two sounds if no known language makes a distinction between them, a property known as "selectiveness". Among the symbols of the IPA, 107 letters represent consonants and vowels, 31 diacritics are used to modify these, and 19 additional signs indicate suprasegmental qualities such as length, tone, stress, and intonation.

IPA is the most convenient notation for the purpose of creating a multilingual speech recognizer because it is applicable to any language and as such it can provide us with the unified phoneme set for all the languages used in the multilingual speech recognizer.

The *Speech Assessment Methods Phonetic Alphabet (SAMPA)*[4] is a computer-readable phonetic script using 7-bit printable ASCII characters, based on the IPA. It was originally developed in the late 1980s for six European languages by the EEC ESPRIT program. As many symbols as possible were taken over from the IPA; where this was not possible, other available symbols were used.

4. PHONEME SET FOR A MULTILINGUAL SYSTEM

Transcription dictionaries for the languages used in the experiments have their own phoneme sets and are described using different notations. These notations are sometimes based on some version of the IPA but more often on the linguistic traditions of the languages. For building a multilingual phoneme set all the phoneme sets of the languages that are to be used have to be reduced to a common denominator.

First, all the dictionaries for all the languages were mapped to SAMPA notation. Decisions on choosing appropriate symbols were based on the description of the notation systems of the given dictionaries, on the information about the phonetic systems of the given languages and on listening to the data. The resulting phoneme set contained 122 phonemes. Error rates for each language trained separately with this phoneme set are shown in Table 3, column IPA1, and error rates for each language in a multilingual system using this phoneme set are shown in Table 3, column MLIPA1.

Then, the number of phonemes was continuously decreased to reduce number of phonemes appearing in one language only. For this purpose, several steps have been taken:

1) All vowels with tones were mapped to corresponding vowels without tones

2) Stressed vowels in Spanish and Portuguese were mapped to corresponding unstressed vowels

3) Nasalized vowels in Spanish and Portuguese were mapped to corresponding unnasalised vowels

4) Long vowels and consonants in all languages were mapped to corresponding short phonemes

5) Phonemes with very little occurrence were mapped to the closest phonemes

The result is a multilingual phoneme set of 93 phonemes, better suited to train a multilingual system. Error rates for each language trained separately with this phoneme set are shown in Table 3, column IPA2, and error rates for each language in a multilingual system using this phoneme set are shown in Table 3, column MLIPA2.

Table 3 shows, that the error rate increases mostly because of training different languages together in a multilingual system (compare columns IPA1 and MLIPA1, IPA2 and MLIPA2), not because of the reduced phoneme set (compare columns Baseline, IPA1 and IPA2). For Vietnamese, for example, the phoneme set IPA1 is 5 times smaller than the baseline due to the elimination of tones, but the error rate is only slightly higher.

[3] http://www.langsci.ucl.ac.uk/ipa/

[4] http://www.phon.ucl.ac.uk/home/sampa/

Vers.	Baseline	IPA1	MLIPA1	IPA2	MLIPA2	S-auto
CZ	24.6	25	27.3	30.6	29.8	30.2
EN	17.6	17.8	24	18	24	25.6
GE	35.8	36	39.7	35.6	44.3	45.9
PO	28	28.7	36.9	31.3	38.3	39.1
RU	35.1	35.2	38.7	35.8	40.5	42.6
SP	29.7	29.8	31.4	29.6	35.2	36.2
TU	34.3	34.4	38.9	34.4	39.2	39.7
VI	28.5	30.2	37.4	30.3	37.7	39.1

Table 3. Baseline results; manual mapping: 1) IPA mapping with 122 phonemes, 2) multilingual system with 122 phonemes, 3) IPA mapping with 93 phonemes, 4) multilingual system with 93 phonemes; semi-automatized mapping: multilingual system

5. UNKNOWN LANGUAGE SPEECH RECOGNITION

One of the possible uses of a multilingual system is speech recognition for languages with no training data. For the following experiments, all the data for seven languages is combined to train a multilingual system. The eighth language is a test language, on which speech recognition is done. For this language, there is only the test data and the proninciation dictionary. Different approaches of constructing a multilingual phoneme set influences the efficiency of speech recognition in this setting.

5.1. Manual mapping

With fully manual mapping the error rates (Table 4, column Manual(93phn)) were too high in most of the cases. Some languages can derive information from other languages, as Czech, for example, may get a lot of information from Russian, and Spanish and Portuguese add to each other. But the overall results are still not satisfactory.

5.2. Semi-automatized mapping

To improve the results, another approach was tested, which makes use of the phoneme confusion matrix, obtained from Multi-lingual Neural Network, similar to what we have used in [10]. In our case it was a perceptron with 1 hidden layer and separate output layers for each language (see Fig. 1). Like this, similar phonemes from different languages are not put in direct competition during the training, since they belong to different softmaxes. On the other hand, we can still generate the posteriors of all the output layers and see the decision. Often, we will see that there is a phoneme match across the languages, as it is demonstrated in Fig. 2. We have used this property to construct an inter-phoneme similarity measure based on Multilingual confusion matrix. This matrix is accumulated by adding the 8-language compound posteriors to a row, the row number is given by the phoneme in the phonetic alignment. Finally, the matrix rows are re-normalized by the numbers of summands. The matrix is shown in Fig. 3, where each matrix element is the average posterior probability of phoneme (the column) given the phoneme in the annotation (the row), which is our similarity measure. Note the secondary diagonals, which show that there are lots of pairs with high similarity across the languages. From this matrix, the most frequent confusions were taken to define the merging of phonemes for further mappings.

Most of the confusion cases were predictable and corresponding mergings have already been made for manual mapping. For example, tones in Vietnamese, stressed and unstressed vowels in Spanish and Portuguese were among the most confused. Some of the confusion cases were of no interest to us, as we do not intend to merge phonemes inside one language: for example, we are not interested in merging /θ/ and /s/ in Spanish, as those phonemes can be sense-distinctive. Most of the information gained from the confusion table is the information about the vowels. Pronunciation of vowels is very variable, much more than the pronunciation of the consonants, so confusion statistics can help choose between two variants of vowel mapping which seem equally plausible, e.g. /ɪ/ and /i/, /ʊ/ and /u/.

The resulting multilingual phoneme set contains 80 phonemes. In the multilingual environment, this phoneme set shows 1-3% increase of error rate for different languages comparing with manual approach (see Table 3, column Semi-auto), but for the unknown language case, for some languages the error rate decreases dramatically - see Table 4, column Semi-auto(80phn).

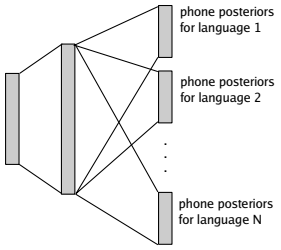

Fig. 1. Multi-lingual neural network

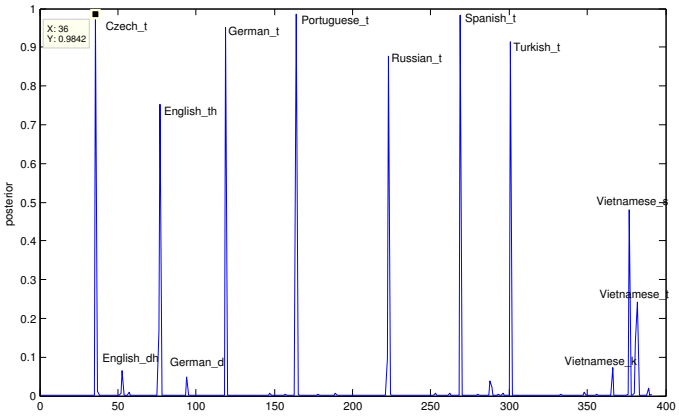

Fig. 2. Example of posteriors of different phonemes across different languages

Phnset	Manual(93phn)	Semi-auto(80phn)	Best
CZ	70.1	80.9	61.3
EN	92.9	94	78.1
GE	92.9	91.2	90.4
PO	90	78.1	65.3
RU	96.9	77.5	75
SP	85.8	60.1	60.1
TU	85.8	72.9	72.9
VI	95.2	93.7	92.8

Table 4. Unknown language speech recognition with different phoneme sets

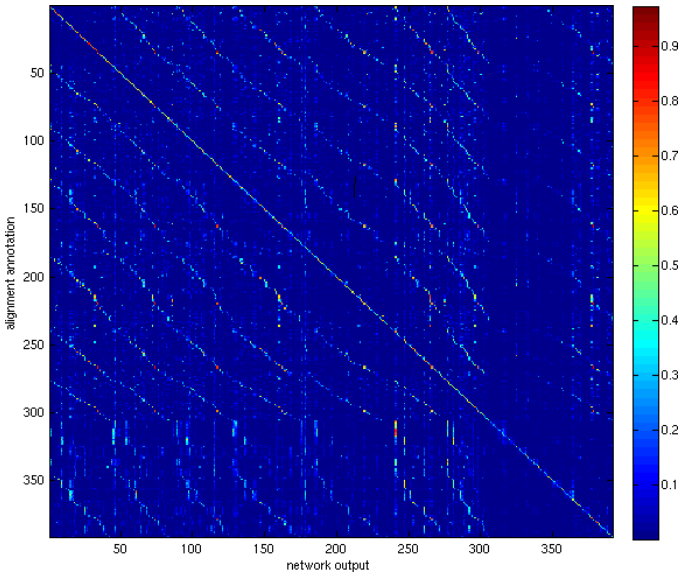

Fig. 3. Multi-lingual confusion matrix

This decrease was caused by merging more phonemes which appear only in one language to the phonemes that appear in several languages. For example, the biggest change in Russian was caused by mapping palatalized consonants, which are characteristic to Russian language to corresponding non-palatalized phonemes. This change is not plausible from the phonetic point of view, as the characteristics of those consonants are very different, but the confusion matrix showed, that in automatic recognition they are close enough to merge them. It helps to get at least some information for those phonemes from the multilingual acoustic models. However, for some languages, such as Czech and English, this new mapping caused increase of error rate, due to more aggressive vowel merging and lower number of phonemes in general.

5.3. Best version for unknown language recognition

For both manually and semi-automatically constructed phoneme sets, a test language usually contains a couple of phonemes which do not occur in the 7-language multilingual system. As there are no acoustic models for these phonemes, the words containinig these phonemes are just skipped during the construction of the recognition network, which yields high error rate. To solve this problem, the best of the two phoneme sets (manual and semi-automatic) was chosen for every test language, and the phonemes, which occur only in test language and are not represented in the 7-language multilingual phoneme set, were mapped to the closest phoneme which appears in one of the 7 training languages. This helps to extract at least some information for these phonemes, even though the phonemes merged are not very close.

The main drawback of this tuning is that the mapping is done for each language individually and manually, so there is no resulting multilingual phoneme set. However, it helps further decrease the error rate by 1-14 % (see Table 4, column Best).

5.4. Systems with 1 hour and 20 minutes of target language

Further experiments were made on 1 hour or 20 minutes of target language speech trained together with all the data for the other 7 languages and on just 1 hour or 20 minutes of target language separately to see if the error rate is lower with the addition of the multilingual data (see Table 5). In one hour of target language setting, only Portuguese shows decrease of error rate when the data of the other languages is added. This is probably due to the fact that the languages chosen are very different in their phonetic systems, whereas Portuguese retrieves a lot of information from Spanish.

System	1hr	ML+1hr.	20min	ML+20min
CZ	49	54.5	60.2	**59.1**
EN	33.8	57.2	52.1	68.2
GE	61	61.8	69.1	70.9
PO	58.8	**56.4**	69.5	**58.6**
RU	63.2	66.4	71.7	71.8
SP	50.7	51.9	59.3	**54.1**
TU	60.5	64.7	68.6	68.8
VI	55	77.5	76.7	87

Table 5. 1 hour and 20 minutes systems with and without the multilingual system

On the 20 minutes of target language data, the results are slightly better with the addition of another languages also for Czech and Spanish, but generally for so many different languages even 20 minutes of target language is better trained alone, at least in such a simple setting.

6. CONCLUSION

A comparison was made between manual and semi-automatic approaches to building a multilingual phoneme set. The two approaches were compared in cases of 1) a multilingual system with abundant data for all the languages, 2) multilingual systems excluding target language 3) multilingual systems with small amount of data for target languages. The work shows that careful choice of merging methods can help improve recognition of languages with no or little training data and reasonably reduce multilingual phoneme set without losing a lot of accuracy.

7. REFERENCES

[1] A. Constantinescu and G. Chollet, "On cross-language experiments and data-driven units for alisp (automatic language independent speech processing)," in *Proc. ASRU 1997*, dec 1997, pp. 606 –613.

[2] B. Wheatley, K. Kondo, W. Anderson, and Y. Muthusamy, "An evaluation of cross-language adaptation for rapid hmm development in a new language," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1, pp. 237–240, 1994.

[3] V. N. Thang, S. Tim, K. Franziska, and T. Schultz, "Rapid bootstrapping of five eastern european languages using the rapid language adaptation toolkit," in *Proceeding of Interspeech*, 2010, pp. 865–868.

[4] H. Lin, L. Deng, D. Yu, Y. Gong, A. Acero, and C.-H. Lee, "A study on multilingual acoustic modeling for large vocabulary asr," in *ICASSP*, 2009, pp. 4333–4336.

[5] U. Bub, J. Kohler, and B. Imperl, "In-service adaptation of multilingual hidden-markov-models," in *Proc. ICASSP 1997*. IEEE Signal Processing Society, 1997, pp. 1451–1454.

[6] J. Köhler, "Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1, may 1998, pp. 417 –420 vol.1.

[7] S. Witt and S. Young, "Language learning based on non-native speech recognition," in *In Proceedings of Eurospeech*, 1997, pp. 633–636.

[8] T. Schultz, M. Westphal, and A. Waibel, "The globalphone project: Multilingual lvcsr with janus-3," in *in Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop, Plzen, Czech Republic*, 1997, pp. 20–27.

[9] F. Grézl, M. Karafiát, and M. Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Proc. ASRU 2011*, dec 2011, pp. 359–364.

[10] K. Vesely, M. Karafiat, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *In: Proceedings of IEEE 2012 Workshop on Spoken Language Technology, Miami, US, IEEESP*, 2012, pp. 336–341.