The COST 278 MASPER initiative - crosslingual speech recognition with large telephone databases

Andrej Žgank[*1], Zdravko Kačič[*1], Frank Diehl[◇2], Klara Vicsi[⋆], Gyorgy Szaszak[⋆], Jozef Juhar[○], Slavomir Lihan[○]

[*] University of Maribor, Maribor, Slovenia
[◇] Universitat Politecnica de Catalunya, Barcelona, Spain
[⋆] Budapest University of Technology and Economics, Budapest, Hungary
[○] Technical University of Košice, Košice, Slovakia

Abstract

This paper presents the work on crosslingual speech recognition carried out by the MASPER initiative that was formed as a part of the COST 278 Action. Two different approaches for transfering monolingual source acoustic models to a new language were compared. The first one was expert-driven, based on the IPA scheme. The second was data-driven, based on a crosslingual phoneme confusion matrix. German, Spanish, Hungarian and Slovak were used as sourcelanguages. Slovenian was selected to be the target language. All experiments were carried out on SpeechDat databases. The results' analysis showed that the expert-driven method outperforms the data-driven one, and that similarities between source and target language have a significant influence on the performance.

1. Introduction

The MASPER initiative (http://masper.uni-mb.si) was established inside the European COST 278 Action "Spoken Language Interaction in Telecommunications" and addresses the multilingual and crosslingual speech recognition. This research topic is very important, as Europe is a multi-cultural society with many languages used in parallel. In the last years more and more speech recognition systems migrate from laboratory environment into real life applications increasing the demand for speech databases in different languages. As the costs of generating a non-existing speech database can be very high, one possibility is to use crosslingual speech recognition. The idea behind this is to transfer the existing source acoustic models from one language to a target language without using speech database in this target language.

Similarity measures used to transfer the source acoustic models to a target language can be divided into two major groups (Schultz, 2000; Žgank, 2003). Expert-driven methods form the first group. Mapping is performed using human knowledge and is usually based on some acoustic-phonetic characteristics. One of the most frequently used approaches is the use of the IPA scheme (IPA Homepage). Expert knowledge from all included languages is needed by such an approach, which can be very difficult if a large number of languages is included. Also, some subjective influence from the mapping can be expected in experiments.

The second group of crosslingual speech recognition approaches is based on data-driven similarity measures. In this case, the similarity measure applied during mapping is calculated from some data. One frequently used method is based on a phoneme confusion matrix. Almost none expert knowledge is needed. The disadvantage of applying this method is that some amount of speech material in the target language is needed to determine the similarity between source and target language acoustic models. This amount of speech material is much smaller (less than 10%) as a complete speech database.

In this paper two different similarities measures for crosslingual speech recognition are compared. Also the influence of language similarities are analyzed. All scripts and procedures developed in the framework of MASPER are publicly available and can be acquired from the MASPER homepage.

2. Crosslingual algorithms

The main influence on the performance of crosslingual speech recognition is the method applied for the transfer of the source acoustic models to the target language. Two different approaches were employed to determine the similarity between the incorporated data and the target language.

2.1. Expert-driven case with IPA scheme

The first approach is based on the IPA scheme (IPA Homepage; Schultz, 2000) which defines acoustic-phonetic properties for phonemes in all languages. For each target language phoneme an equivalent phoneme in the source language was searched for. As an equivalent phoneme, the source phoneme with the same IPA symbol was selected. The ratio of equivalent phonemes depends on the similarity of languages and on the number of phonemes in each involved language. In case of the IPA equivalent was non-existent, the most similar phoneme according to the IPA scheme was looked for. The search for the most similar candidate can be performed in horizontal or vertical direction through the IPA scheme. The main advantage of the IPA method is that it could be applied without any speech material in the target language. On the other side, expert knowledge is needed and also subjective influence is introduced by the expert. The main reason for this is that the same IPA symbol can be pronounced slightly different between languages.

[1]This work was supported by the Slovenian Ministry of Education, Science and Sport under contract number PP-0796/99.

[2]This work was granted by CICYT under contract number TIC2002-04447-C02.

As the phoneme set in the SpeechDat databases consist of SAMPA phonemes (SAMPA Homepage) that are computer readable representation of IPA symbols, SAMPA symbols could be used directly in the experiments instead of converting the phoneme set to IPA and vice-versa.

2.2. Data-driven case with phonetic confusion matrix

The second approach for crosslingual speech recognition introduced in the acoustic modelling phase was the one based on a phoneme confusion matrix (Schultz, 2000; Žgank, 2003). The idea behind this method is that similar phonemes are confused during speech recognition with a phoneme recogniser. The characteristic of such a speech recogniser is that it recognises phoneme sequences instead of words. For generating crosslingual confusion matrix, acoustic models of one of the source languages were applied on speech utterances of the target language. The recognised sequence of source phonemes was then aligned to the reference sequence of the target phonemes. The output of this alignment was the crosslingual phoneme confusion matrix M. Now, for each target phoneme ϕ_{trg}, the best corresponding source phoneme ϕ_{src} was searched for. As similarity measure, the number of phoneme confusions $c(\phi_{trg}, \phi_{src})$ was selected. The target phoneme ϕ_{trg} was defined as:

$$\phi_{trg} = max\ c(\phi_{trg}, \phi_{src}) \quad (1)$$

For each target phoneme ϕ_{trg} the source phoneme ϕ_{src} with the highest number of confusions c was determined. If two or more source phonemes had the same highest number of confusions c, the decision which one should represent the target phoneme ϕ_{trg} was left over to the expert. The same procedure was employed in case of no confusions between target and source phonemes.

The advantage of a crosslingual similarity measure based on a confusion matrix is that it is fully data-driven and almost no expert knowledge is needed. The disadvantage is that target language speech material must be available to be able to generate the phoneme confusion matrix, but already a small amount is sufficient.

3. SpeechDat databases

The speech recognition experiments were performed using different SpeechDat fixed telephone databases (Hoege et al., 1997; van den Heuvel et al., 2001/1). The SpeechDat project was initialized in the year 1996 and covers at the moment more than 50 languages. All databases were generated according to the same standard and have identical structure. The objectives of SpeechDat are voice driven telephone applications. The number of speakers per language varies between 500 and 5000 and depends on the population size per language. For each speaker 43 different utterances were recorded (van den Heuvel et al., 2001/2).

As source languages, the following SpeechDat databases were present:

- German 4000 FDB SpeechDat(II) - DE,
- Spanish 4000 FDB SpeechDat(II) - ES,
- Hungarian 1000 FDB SpeechDat(E) - HU,
- Slovak 1000 FDB SpeechDat(E) - SK.

To have the same number of speakers per language only 1000 speakers were selected from the German and Spanish database. As target language Slovenian 1000 FDB SpeechDat(II) - SI (Kaiser, 1998) was applied. Sentences inappropriate for speech recognition (van den Heuvel et al., 2001/2), were excluded from the training set. After this, the training set for each source language consisted of approximately 30.000 utterances.

As the acoustic channel is very important in case of crosslingual speech recognition, first signal to noise ratio (SNR) for all included databases was calculated and analyzed. All utterances were grouped according to being in the training or test set. Then they were classified into five different categories, as can be seen in Table 1.

Comparing the training set SNR values, it can be seen that the noise level is the lowest in German database, followed by the Spanish database. The highest noise level was found in the Slovenian training set, where 76.0% of the phrases have SNR between 10 and 20 dB. Similar conditions can be observed for the test set. Again the German part has the lowest noise level followed by Spanish. For the Slovenian test set the ratio of utterances with SNR between 10 and 20 dB increases to 81.7%.

4. Monolingual setup

Monolingual speech recognisers were needed for two different purposes. The first goal was to build source acoustic models that were then applied in crosslingual experiment. The second goal was to evaluate a pure monolingual Slovenian speech recogniser that served as reference for crosslingual experiments.

For the monolingual MASPER1 script, the refrec0.96 system (Lindberg et al., 2000; Johansen et al., 2000) was used as a starting point. Different modifications were necessary to it suitable for crosslingual environment. All training procedures were fully language independent, with some configuration files (e.g mapping of rare phonemes, phonetic broad classes) that contained language specific information. In the first step, context independent acoustic models were generated, and then the context dependent ones were built. As they can assure better quality, the context dependent acoustic models are more suitable to be used for source acoustic models. Due to the larger number of models they provide a better coverage of the acoustic feature space. More details about the training procedure can be found in (MASPER Homepage) and (Lindberg et al., 2000)

All speech recognisers were evaluated on six different test scenarios (van den Heuvel et al., 2001/2):

- Application words (A),
- Yes/no answers (Q),
- Isolated digits (I),
- Connected digits (B),
- City names (O),
- Phonetically balanced words (W).

Language	0-10dB	10-20dB	20-30dB	30-40dB	40-50dB	50-60db	60dB<
DE-TRN	0.0	0.2	5.9	34.0	47.0	12.2	0.6
ES-TRN	0.0	0.7	10.4	42.5	40.1	5.7	0.5
HU-TRN	0.0	1.1	13.5	48.5	29.5	6.0	1.6
SK-TRN	0.1	2.2	16.1	43.7	34.6	3.4	0.0
SI-TRN	0.7	8.4	76.0	14.9	0.0	0.0	0.0
DE-TST	0.0	0.1	6.5	37.9	42.4	10.5	2.5
ES-TST	0.1	1.5	11.6	43.2	34.9	6.9	1.8
HU-TST	0.0	0.6	13.4	46.5	30.9	7.0	1.5
SK-TST	0.6	2.4	19.0	48.6	28.3	1.1	0.0
SI-TST	0.2	12.0	81.7	6.1	0.0	0.0	0.0

Table 1: SNR distribution over the training (TRN) and the test (TST) set. All numbers in percentage of the subset.

The simplest test set had only 2 words in the vocabulary, the hardest one several thousand words.

5. Mapping to the target language

The phoneme mapping pairs between source and target phonemes produced by both methods were compared and analyzed. The Slovenian target phoneme set consisted of 39 different phonemes, some rare phonemes were mapped into more frequent ones. In the Table 2, the number of phonemes in the source languages and the ratio of mapping pairs with the same SAMPA symbol – using the IPA scheme (IPA) and the data-driven confusion matrix (CM) – can be seen.

Language	Set size	IPA (%)	CM (%)
DE	47	76.9	33.3
ES	30	53.9	30.8
HU	68	76.9	41.0
SK	51	74.4	56.4

Table 2: Number of phonemes in each source language and ratio of phoneme pairs per language with the same symbol.

The overall highest number of overlapping Slovenian phonemes was achieved for the Slovak language which belongs to the same language group. The overlap for Hungarian and German is higher than for Slovak in case of the IPA method, but significantly lower than in the data-driven case. The lowest overlap occured for Spanish as it has also the smallest phoneme set of all involved languages. When a phoneme confusion matrix was applied, it often occured that the target and source phonemes differed only in length, which is also reflected in a lower overlap ratio for the data-driven case.

For performing the speech recognition experiment in target language, context dependent source acoustic models with 32 Gaussian probability density functions per state were employed, as they tend to offer better results. Using the source-target mapping pairs, the target language triphone set was converted into each of the source languages. As some of these converted source triphones were non-existent, a phonetic decision tree based clustering approach was used to find existing counterparts for the unseen triphones.

6. Speech recognition results

6.1. The monolingual case

For all languages (source and target) included in the experiment, monolingual acoustic models were built and tested. The word error rates (WER) for six different test scenarios are presented in Table 3.

Language	A	Q	I	BC	O	W
DE	1.70	0.00	0.00	1.11	3.30	7.03
ES	1.81	1.19	3.68	3.12	15.96	8.15
HU	0.17	0.57	0.00	0.78	3.59	5.53
SK	0.43	0.00	0.00	1.17	7.90	10.46
SI	2.15	0.58	4.66	2.46	6.19	13.48

Table 3: Word error rate for monolingual speech recognition.

From the results in Table 3 it can be seen that the WER inside one language mostly depends on the size of the vocabulary. The best results were achieved with the simplest test sets - yes/no answers and digits. For German, Hungarian and Slovak, the WER in some cases was 0.00%. Difficult test sets in all languages were the O and W. For these cases, the WER was higher than for other sets. Usually the worst result was for phonetically balanced words. The worst result of 15.96% WER was achieved for the O test set in Spanish language.

6.2. Crosslingual case

In the final step of the experiments the comparison of speech recognition results for both crosslingual speech recognition methods was performed. Transferred source acoustic models from all source languages were now used for Slovenian (target) speech recognition. First, the results of the IPA scheme approach are presented in Table 4.

The results of the IPA scheme method shows that the best result was always achieved with Slovak source acoustic models. The difference between crosslingual acoustic models and Slovenian monolingual reference acoustic models (row 6 in Table 3) was smaller for simpler test sets. For example 1.16% versus 0.58% for the Q test set, which is also the smallest difference between the crosslingual and the reference system. With increased complexity of the recognition test also the performance of crosslingual acoustic mod-

Source	A	Q	I	BC	O	W
DE	31.12	6.07	36.27	33.89	47.42	61.68
ES	60.37	5.78	68.91	79.50	81.96	89.59
HU	27.01	5.20	20.21	25.93	37.63	52.60
SK	18.88	1.16	17.10	22.46	27.32	39.65

Table 4: Slovenian crosslingual speech recognition performance with IPA scheme approach.

els dropped for all languages. The hardest task was always the W test set.

The second best source acoustic models for Slovenian speech recognition where those from the Hungarian language being followed by the German acoustic models. With except of the Q test set the worst results were obtained with Spanish source acoustic models. The overall worst result was for the W test set with 89.59% WER.

The speech recognition results for the second method based on crosslingual phoneme confusion matrix approach are presented in Table 5.

Source	A	Q	I	BC	O	W
DE	40.75	6.36	32.12	30.87	69.59	73.16
ES	76.64	5.78	74.09	78.56	96.91	97.46
HU	31.87	5.20	25.91	30.60	47.42	69.43
SK	19.81	0.58	13.99	16.59	32.99	40.19

Table 5: Slovenian crosslingual speech recognition performance based on crosslingual phoneme confusion matrix approach.

As with the IPA method, the best result was achieved with the Slovak source acoustic models. With the Q test set, the results were the same as when Slovenian monolingual reference acoustic models were applied. Again, when the test complexity increased, the difference between crosslingual acoustic models and reference system growed. When Hungarian and German acoustic models were used as source, the obtained results where worse than for Slovak language. The worst results were produced with Spanish source acoustic models. This was already indicated by Spanish / Slovenian phoneme confusion matrix, where different Slovenian phonemes were mapped into the same Spanish phoneme. A possible explanation would be that Spanish and Slovenian language are just to different. Also the small number of Spanish phonemes has an important influence on this topic.

If the IPA and the phoneme confusion matrix method are compared, it can be seen that the IPA scheme method tends to give better speech recognition results. In some cases when similar source language was applied, the data-driven method produced better results. A probable explanation is that in such cases the crosslingual phoneme recognition generated better confusion matrix as source and target phoneme were similar.

From the presented results it can be also concluded that source and target language similarity plays an important role. The best Slovenian crosslingual speech recognition results were always achieved with Slovak source acoustic models that belong to the same Slavic language group.

7. Conclusions and future work

This paper presented work done in the scope of the COST 278 MASPER special interest group. A expert-driven and a data-driven method for crosslingual speech recognition were compared. The achieved results showed that in case of monolingual source acoustic models the expert-driven method outperforms the data-driven approach. A further conclusion which could be made from the results' analysis is that language similarity has significant influence on speech recognition results. In the future also multilingual source acoustic models will be incorporated in the language set.

8. References

Henk van den Heuvel, Jerome Boudy, Zsolt Bakcsi, Jan Cernocky, Valery Galunov, Julia Kochanina, Wojciech Majewski, Petr Pollak, Milan Rusko, Jerzy Sadowski, Piotr Staroniewicz, Herbert S. Tropf, 2001. SpeechDat-E: Five Eastern European Speech Databases for Voice-Operated Teleservices Completed. *Proc. Eurospeech 2001*, Aalborg, Denmark.

van den Heuvel, H., Boves, L., Moreno, A., Omologo, M., Richard, G., Sanders, E., 2001. Annotation in the SpeechDat Projects. *International Journal of Speech Technology*, 4(2):127 – 143.

Hoege, H., H. Tropf, R. Winski, H. van den Heuvel, R. Haeb-Umbach 1997. European speech databases for telephone applications. *Proc. ICASSP '97*, Muenchen.

IPA Homepage, http://www2.arts.gla.ac.uk/IPA/ipa.html.

Johansen, F.T., Warakagoda, N., Lindberg, B., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., Salvi. G., 2000. The COST 249 SpeechDat Multilingual Reference Recogniser. *Proc. LREC'2000, Athens*, Greece.

Kaiser, J., Z. Kačič, 1998. Development of the Slovenian SpeechDat database. *Proc. Speech Database Development for Central and Eastern European Languages*, Granada.

Lindberg, B., Johansen, F.T., Warakagoda, N., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., Salvi, G., 2000. A noise robust multilingual reference recogniser based on SpeechDat(II). *Proc. ICSLP 2000, Beijing*, China.

MASPER Homepage, http://masper.uni-mb.si.

SAMPA Homepage http://www.phon.ucl.ac.uk/home/-sampa/home.htm

Schultz, T., 2000. *Multilinguale Spracherkennung - Kombination akustischer Modelle zur Portierung auf neue Sprachen*. Ph.D. thesis, University of Karlsruhe, Germany.

Žgank, A., 2003. *Data driven method for the transfer of source multilingual acoustic models to a new language*. Ph.D. thesis, University of Maribor, Slovenia.