Crosslingual Speech Recognition with Multilingual Acoustic Models Based on Agglomerative and Tree-Based Triphone Clustering

Andrej Žgank[1], *Bojan Imperl*[1], *Finn Tore Johansen*[2],
Zdravko Kačič[1], *Bogomir Horvat*[1]

[1]Faculty of Electrical Engineering and
Computer Science, University of Maribor,
Smetanova 17, SI-2000 Maribor, Slovenia
andrej.zgank@uni-mb.si

[2]Telenor Research and Development
P.O. Box 83, N-2007 Kjeller, Norway
finn-tore.johansen@telenor.com

Abstract

The paper describes our ongoing work on crosslingual speech recognition based on multilingual triphone hidden Markov models. Multilingual acoustic models were built using two different clustering procedures: agglomerative triphone clustering and tree-based triphone clustering. The agglomerative clustering procedure is based on measuring the similarity of triphones on a phoneme level where the monophone similarity is estimated by the Houtgast algorithm. The tree-based clustering procedure is based on common broad classes. The Slovenian, German and Spanish 1000 FDB SpeechDat(II) databases were used for training. The crosslingual speech recognition was performed on the Norwegian 1000 FDB SpeechDat(II) database. No adaptation or training with the Norwegian database was used. The mapping of Norwegian phonemes was done with the IPA scheme. Five different Norwegian recognition vocabularies were generated. The best crosslingual system achieved a recognition rate of 45.03%, while the reference Norwegian system achieved 78.32%.

1. Introduction

Globalization of speech technology brings forward the problem of dealing with unknown languages, that is languages with little or no language resources. To be able to develop speech recognition systems for a new language an existing speech database for this language is the necessary prerequisite. The task of creating a speech database for a new language is often very time consuming and expensive. One way to avoid this problem is the possibility of porting an existing system to new language - the crosslingual transfer of speech recognition system.

The purpose of the work presented in this paper was to evaluate the performance of two different triphone based multilingual speech recognition systems in a crosslingual experiment, that is, speech recognition for a new unseen language. Our crosslingual experiments are based on SpeechDat(II) telephone databases [1]. Two sets of multilingual triphone models were created. The first multilingual system was based on an agglomerative triphone clustering technique [2] and the second one on a tree-based triphone clustering [3]. The phonemes from recognition vocabulary of the new language were mapped to existing phonemes of the multilingual models. No training or adaptation to the new language was carried out. Because of this, the recognition rate of the crosslingual speech recogniser was lower. The crosslingual tests with mapped recognition vocabulary were performed on both multilingual speech recognition systems. These experiments present the latest advances in our ongoing work in the area of multilingual speech recognition [4].

The mapping of phonemes from a new language to existing phonemes in a multilingual set can be done following the IPA scheme [5], using a data-driven approach or with a combination of both approaches. Both different methods for mapping phonemes from a new language were used in previous experiments of speech recognition reported by other authors. In [6, 7] the mapping was created with the use of IPA scheme. The data-driven approach was used in [8]. In our experiments the IPA scheme was applied.

The paper is organized as follows. Both clustering procedures and distance measures are presented in Section 2. The databases used are described in Section 3. The experimental systems are presented in Section 4. The crosslingual phoneme mapping is described in Section 5 and the crosslingual recognition results in Section 6. The conclusion is made in Section 7.

2. Clustering procedures and distance measures

Like in monolingual systems the context dependent acoustic models outperform the context independent models also in multilingual systems [9]. In developing a multilingual system for crosslingual speech recognition, two issues should be considered: the recognition rate and the complexity of the recogniser. Systems with lower complexity usually also has lower recognition rate. Our first multilingual system is built using the agglomerative (bottom-up) clustering approach [2] and the second one with the tree-based clustering (top-down) approach [4].

2.1. Agglomerative clustering

The distance measure applied in the agglomerative approach [2] is based on measuring the similarity of triphones on a phoneme level. The similarity between two different triphones φ_i^L-φ_i+φ_i^R and φ_j^L-φ_j+φ_j^R can be estimated by measuring the similarity of the left phoneme φ^L, central phoneme φ and right phoneme φ^R:

$$\begin{aligned} S_{TRI}(\varphi_i^L\text{-}\varphi_i\text{+}\varphi_i^R, \varphi_j^L\text{-}\varphi_j\text{+}\varphi_j^R) = \\ = W_L\ S(\varphi_i^L, \varphi_j^L) + W_C\ S(\varphi_i, \varphi_j) + \\ +W_R\ S(\varphi_i^R, \varphi_j^R), \end{aligned} \tag{1}$$

where S denotes the similarity between two phonemes, W_L, W_C and W_R are the weights for setting the influence of each phoneme-level similarity estimates, and $S_{TRI}(\varphi_i^L\text{-}\varphi_i\text{+}\varphi_i^R, \varphi_j^L\text{-}\varphi_j\text{+}\varphi_j^R)$ is the resulting similarity estimation of both triphones.

The similarity of two triphones in (1) can be based on any phoneme distance measure. We have decided to apply the phoneme distance measure suggested in [10], which is based on a monophone confusion matrix:

$$\begin{aligned} S(\varphi_i, \varphi_j) &= \frac{1}{2}\sum_{k=1}^{N}[\,c(\varphi_i,\varphi_k) + c(\varphi_j,\varphi_k) \\ &\qquad - |c(\varphi_i,\varphi_k) - c(\varphi_j,\varphi_k)|\,] \\ (\varphi_i,\varphi_j) \in \varphi \quad &, \quad 1 \le i,j \le N \;,\; i \ne j, \end{aligned} \tag{2}$$

where $S(\varphi_i, \varphi_j)$ denotes the similarity of two phonemes φ_i and φ_j, N is the number of phonemes, $c(\varphi_i, \varphi_k)$ is the number of confusions between phoneme φ_i and phoneme φ_k. The phoneme φ_i should not be the same as φ_j.

The advantage of the distance measure given by (1) is that it can also handle unseen triphones. Each unseen triphone is compared to all the existing triphones with the use of distance measure and then tied to the most similar one. The problem of unseen triphones is very important in the case of crosslingual speech recognition. Mapping the recognition vocabulary of a new language generates a number of new - unseen triphones.

The triphones that are similar enough in sense of the distance measure can be merged. If an average distance among all triphones from the group is less than a predefined threshold T, the group is equated. The average distance between M triphones is calculated as:

$$\begin{aligned} S_M(\hat{\Theta}) &= \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} S_M(\Theta_i,\Theta_j)}{\sum_{i=1}^{N-1} i} \\ (\Theta_i,\Theta_j) &\in \hat{\Theta} \;,\; 1 \le i,j \le N \;,\; i \ne j, \end{aligned} \tag{3}$$

where Θ_i denotes the triphone $\varphi_i^L\text{-}\varphi_i\text{+}\varphi_i^R$ and Θ_j denotes the triphone $\varphi_j^L\text{-}\varphi_j\text{+}\varphi_j^R$, $\hat{\Theta}$ is the group of triphones and $S_M(\hat{\Theta})$ is the average distance among all triphones from the group $\hat{\Theta}$.

The result of the clustering algorithm is a list of triphones from all languages that can be tied together. Some triphones that are language specific remain untied.

2.2. Tree-based clustering

The second set of multilingual triphone models was generated with the use of a phonetic tree-based clustering procedure, suggested by [3]. To successfully build multilingual acoustic models, common broad classes for all languages were defined. Each phoneme was tagged with a language specific tag and no specific language questions [6] were added. The similar phonemes from all languages were grouped in common categories.

The questions needed for tree building are generated from broad classes. One binary tree is built for each state of the phoneme. The questions are placed in nodes of the tree. At the beginning, all states are placed in the root node of the tree in one cluster. The node is then split into two, by finding the question which gives the maximum increase of log likelihood for the particular training data set. When the increase is smaller than the threshold, the splitting is stopped.

3. Databases

The tests were performed on fixed telephone speech databases SpeechDat(II) [1]. These databases provide a realistic multilingual environment for development of voice driven teleservices. Characteristics and recording conditions of all databases were equal, which is crucial in case of crosslingual and multilingual speech recognition experiments. The following databases were employed:

- Slovenian (SL) 1000 FDB SpeechDat(II),
- German (DE) 1000 FDB SpeechDat(II),
- Spanish (ES) 1000 FDB SpeechDat(II),
- Norwegian (NO) 1000 FDB SpeechDat(II).

Each database consists of recordings of 1000 speakers. Each speaker is represented with approximately 10 minutes of speech. The training part of each database consists of 800 speakers and the remaining 200 speakers were used for test.

Table 1: *The number of test utterances, phonemes and size of recognition vocabulary for all databases.*

Language	Test.utter.	Phonemes	Rec.vocab
SL	748	49	605
DE	674	48	674
ES	681	31	681
NO	784	45	792

Recordings with mispronunciations and unintelligible speech were excluded from the set. More than 20.000 sentences from each database were used in training part of experiments described in this paper. In all experiments only the W1-W4 corpuses [1], containing phonetically rich words, were applied during the test. Test data for each language are presented in Table 1.

4. Experimental systems

The recognisers applied for crosslingual speech recognition were generated with the use of the script refrec0.9 developed in the framework of the ”SpeechDat task force” within COST 249[1] project [11, 12]. The perl script is created on the base of the HTK toolkit [13] and is an extended version of the tutorial example in the HTK Book. To achieve more robust performance of the system with telephone speech, a different feature extraction frontend module than in refrec09 was applied. The acoustic feature vector consisted of 24 mel-scaled cepstral, 12 Δ - cepstral, 12 $\Delta\Delta$ - cepstral, high pass filtered energy, Δ - energy and $\Delta\Delta$ - energy coefficients. The procedure of maximum likelihood channel adaptation [14] was carried out on feature vectors. The number of elements in the feature vector was reduced to 24 with the use of linear discriminant analysis [14].

First the monolingual speech recognisers were developed. The 3 state left-right hidden Markov model (HMM) topology was employed for acoustic models. To avoid the need for language models, tests with isolated words were performed. The triphone models were built and the number of Gaussian mixtures per state was increased to 32.

The results of monolingual speech recognition with monolingual systems for all 4 languages are presented in Table 2. The

[1] COST 249 - Continuous Speech Recognition over the Telephone, http://www.elis.rug.ac.be:80/ELISgroups/speech/cost249/

Table 2: *Monolingual recognition results for Slovenian (SL), German (DE), Spanish (ES) and Norwegian (NO) language with monolingual triphone models.*

Language	Recognition rate (%)
SL	88.25
DE	92.51
ES	93.91
NO	78.32

Norwegian monolingual reference system achieved a recognition rate of 78.32%. The COST249 Norwegian system [11] with standard HTK mel cepstral frontend and a bigger vocabulary achieved a recognition rate of 65.27% on the same test set. This result shows that the selected frontend did improve the performance. The Norwegian monolingual reference recognition rate is smaller than monolingual recognition rate for other three languages, as was already noticed in [12, 11]. The probable cause is the length and the number of words in the recognition vocabulary. The size of recognition vocabularies is presented in Table 1. The Norwegian monolingual speech recognition system was used as the reference system in the crosslingual test.

With the Slovenian, German and Spanish database two multilingual set of triphone models were designed. Both clustering procedures described in Section 2 were applied. The optimal threshold values [4] for both clustering procedures of multilingual triphone models were derived experimentally. The two sets of multilingual triphone models were used for crosslingual speech recognition with Norwegian recognition vocabularies generated with different mappings, described in Section 5.

5. Crosslingual phoneme mapping

For the crosslingual speech recognition the same test set was applied as for the Norwegian reference system. All Norwegian phonemes were mapped to other languages with the use of the IPA scheme. Each Norwegian phoneme was mapped to the IPA symbol of equivalent phoneme in other languages. If the equivalent phoneme did not exist in the target language, the most similar one was chosen (according to IPA notation). The most problematic was the conversion of Norwegian diphthongs. Also problematic was the mapping of Norwegian vowels to Spanish vowels, because the Norwegian language has 18 vowels and Spanish only 5. The mapping resulted in 5 different recognition vocabularies:

- Norwegian to German (ND),
- Norwegian to Spanish (NE),
- Norwegian to Slovenian (NS),
- Norwegian to Multilingual (NM),
- Norwegian to Parallel (NP).

In the case of NM vocabulary Norwegian phonemes were mapped to the optimal phoneme in any of the three target languages. With this procedure the rules of phonotactic were trespassed. In the Norwegian to Parallel (NP) mapping each word in vocabulary has three pronunciation variants: in Slovenian, in German and in Spanish. In the case of NP system the vocabulary size was 3 time larger, but here the data choose the optimal mapping. As mentioned in Section 2, many new unseen triphones occurred after mapping.

Table 3: *Number of all mapped triphones and the percentage of missing triphones for all five mapping configurations.*

Language	Mapped triphones	Missing triphones (%)
ND	2214	30.67
NE	1606	32.50
NS	2085	44.41
NM	2331	63.58
NP	5905	36.02

As can be seen in Table 3 for the German language, from 2214 triphones there are 30.67% missing. As seen in Table 3, the worst case is the NM vocabulary, where 63.58% triphones were missing. In the agglomerative system unseen triphones were tied to existing ones with the use of a distance measure and in the tree-based system with the use of a phonetic decision tree. From Table 3 we can see that the German language, which belongs to the same Germanic language group as Norwegian, ought to be the most similar language. The Slovenian, which belongs to Slavic group and the Spanish, which belongs to Romanic group are less similar to Norwegian.

6. Crosslingual speech recognition

The crosslingual speech recognition experiments were carried out without any training or adaptation on a Norwegian database. All five mapping configurations of a Norwegian recognition vocabulary, presented in Section 5, were tested with the agglomerative multilingual triphone models and with the tree-based multilingual triphone models. The test results for the crosslingual transfer are presented in Table 4. As can be seen the tree-based multilingual system performed better than the agglomerative multilingual system. The difference between agglomerative and tree-based systems is approximately 10%. From all one language mapping vocabularies (ND, NE, NS) that were used for Norwegian crosslingual speech recognition, the German mapping vocabulary performed best in both systems. With tree-based triphone models it achieved 43.62% recognition rate and 34.06% with agglomerative triphone models. This result was anticipated due to experience from the mapping.

Table 4: *Recognition rate for crosslingual speech recognition with both multilingual triphone systems and five vocabulary mapping configurations.*

Map.config	Agglomerative (%)	Tree-based (%)
ND	34.06	43.62
NE	19.13	34.57
NS	1.91	1.28
NM	25.65	32.53
NP	33.42	45.03

The mapping of the Norwegian recognition vocabulary to Spanish achieved a recognition rate of 34.57% with tree-based multilingual models and 19.13% with agglomerative multilingual models. The lower result for the Spanish mapping was also expected because of the low number of Spanish phonemes, especially the vowels. It is known that vowels are very important for speech recognition. The results of Norwegian speech recognition applying the Slovenian mapping configuration shows, that there are some remaining problems in this configuration, which must be investigated in the future. One possible problem would be in the type of mapping used. Maybe some other mapping configuration would produce better results. The other explanation for such result can hide in the difference between language groups to which both languages belong. Additional tests will be performed in the future to investigate this problem.

In the case of the multilingual vocabulary NM, each triphone can consists of phonemes from different languages. The distance of such triphones to existing ones is in average smaller than for other cases, which shows that the similarity is also smaller. The NM mapping system achieves recognition rates similar to the NE system. This is probably due to the trespassing of phonotactic rules. The best mapping configuration in the tree-based case was when all three mapping possibilities for each word were in the vocabulary (NP). In this case the data choose the best pronunciation variant. The system achieved a recognition rate of 45.03% with tree-based clustered multilingual triphone models. The test showed that in 71.68% the German pronunciation was chosen, in 26.91% the Spanish and only in 1.41% the Slovenian one. The best system with tree-based models and NP mapping vocabulary achieved similar results to comparable systems [8].

7. Conclusion

We have tested crosslingual speech recognition for the Norwegian language with multilingual triphones, without training or adaptation. The best performance was achieved with tree-based clustered multilingual triphone models and a recognition vocabulary with parallel (NP) mapping configuration. The disadvantage of this system is that the complexity of recogniser is 3 time larger than by other mapping configurations. In the future problems with the Slovenian mapping will be thoroughly investigated. More data oriented mapping procedures will be studied. It is known [15] that the IPA scheme is subjective in some cases. The influence of the number of languages in the multilingual set of triphone models should be investigated. It is possible that an increased number of languages would improve the recognition rate, because larger number of different languages would better cover the phonetic inventory of the new language. Also some adaptation or training technique on a small number of Norwegian utterances will be tested.

Acknowledgments

The authors wish to acknowledge Siemens AG and the Universitat Politechnica de Catalunya for providing the German and Spanish SpeechDat databases. The authors also wish to thank Melita Zemljak (Faculty of Education, University of Maribor) for her help with the IPA scheme.

8. References

[1] Höge, H., Tropf, H., Winski, R., van den Heuvel, H., and Haeb-Umbach, R., ”European Speech Databases for Telephone Applications”, Proc. ICASSP’97, pages 1771 – 1774, Munich, Germany, 1997.

[2] Imperl, B. and Horvat, B., ”The Clustering Algorithm for the Definition of Multilingual Set of Context Dependent Speech Models”, Proc. EUROSPEECH’99, Budapest, Hungary, 1999.

[3] Young, S., Odell, J., Woodland, P., ”Tree-based State Tying for High Accuracy Acoustic Modelling”, Proc. ARPA Human Language Technology Conference, Plainsboro, USA, 1994.

[4] Imperl, B., Kačič, Z., Horvat, B., Žgank, A., ”Agglomerative vs. Tree-Based Clustering for the Definition of Multilingual Set of Triphones”, Proc. ICASSP’2000, Istanbul, Turkey, 2000.

[5] ”The IPA 1989 Kiel Convention”, Journal of the International Phonetic Association 1989(19) pages 67 – 82.

[6] Schultz, T. and Waibel, A., ”Multilingual and Crosslingual Speech Recognition”, Proc. DARPA Broadcast News Workshop 1998, Lansdowne, USA, 1998.

[7] Schultz T. and Waibel A., ”Polyphone Decision Tree Specialization for Language Adaptation”, Proc. ICASSP’2000, Istanbul, Turkey, 2000.

[8] Köhler, J., ”Multi-lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds”, Proc. ICSLP’96, Philadelphia, USA, 1996.

[9] Finke, M. and Rogina, I., ”Wide Context Acoustic Modeling in Read vs. Spontaneous Speech”, Proc. ICASSP’97, Munich, Germany, 1997.

[10] Andersen, O., Dalsgaard, P., and Barry, W., ”Data-driven Identification of Poly- and Mono-phonemes for Four European languages”, Proc. EUROSPEECH’93, pages 759 – 762, Berlin, Germany, 1993.

[11] Johansen, F. T., Warakagoda, N., Lindberg, B., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., and Salvi, G., ”The COST 249 SpeechDat Multilingual Reference Recogniser”, XLDB - Very Large Telephone Speech Databases, LREC’2000 Workshop Proc., Athens, Greece, 2000.

[12] Lindberg, B., Johansen, F. T., Warakagoda, N., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., and Salvi, G., ”A Noise Robust Multilingual Reference Recogniser Based on SpeechDat(II)”, Proc. ICSLP’2000, Beijing, China, 2000.

[13] Young, S., ”The HTK Book Version 2.1”, Cambridge University, 1997.

[14] Haunstein, A. and Marschall, E., ”Methods for Improved Speech Recognition Over the Telephone Lines”, Proc. ICASSP’95, Detroit, USA, 1995.

[15] Köhler, J., ”Comparing Three Methods to Create Multilingual Phone Models for Vocabulary Independent Speech Recognition Tasks”, Multi-lingual Interoperability in Speech Technology, Proc. ESCA-NATO Tutorial and Research Workshop, Leusden, Netherlands, 1999.