Conversion from Phoneme Based to Grapheme Based Acoustic Models for Speech Recognition

Andrej Žgank and Zdravko Kačič

Laboratory for Digital Signal Processing,
University of Maribor, Smetanova ul. 17, SI-2000 Maribor, Slovenia
andrej.zgank@uni-mb.si

Abstract

This paper focuses on acoustic modeling in speech recognition. A novel approach how to build grapheme based acoustic models with conversion from existing phoneme based acoustic models is proposed. The grapheme based acoustic models are created as weighted sum from monophone acoustic models. The influence of particular monophone is determined with the phoneme to grapheme confusion matrix. Further, the context-dependent acoustic models are being trained within the grapheme training procedure. The decision tree based clustering approach is used to tie similar states. A modified data-driven method for generation of grapheme broad classes needed during the initialization of decision tree is being applied. The data-driven broad classes are created using the grapheme based confusion matrix. All experiments were performed with the Slovenian language (1000 FDB SpeechDat(II) database), which is a highly inflectional language with no fixed set of rules for grapheme to phoneme conversion. The achieved results showed improvements of speech recognition results with the proposed methods.

Index Terms: acoustic modeling, grapheme based, bootstraping, confusion matrix, speech recognition.

1. Introduction

With modern technological development the automatic speech recognition (ASR) systems are becoming more and more present in everyday live. First real-life applications were mainly based on isolated words recognition, thereafter the systems based on large vocabulary continuous speech recognition (LVCSR) followed. But there are still groups of languages, where there is a lack of methods how to built a speech recogniser with performance comparable to performance of an English LVCSR system (e.g. agglutinative languages, inflectional languages).

One of such languages is Slovenian, which belongs to the group of south Slavic languages. It is a highly inflectional language with a relatively free word order. These language's peculiarities results in a very large number of different words that can be formed from one lemma. Approximately 10 time larger Slovenian vocabulary is needed to ensure the same out of vocabulary rate as for English.

Phonemes are the smallest still distinguishable speech units by humans and are as such usually used in speech recognisers. To be able to link the orthographic transcription of a word with the phoneme based acoustic models, a phonetic vocabulary is needed. A phonetic vocabulary for a speech recognition system can be generated using a large spectrum of approaches, from manual ones to different automatic methods. In case of highly inflectional languages, the huge number of different words makes it unrealistic to manually build a phonetic vocabulary with high word coverage. Slovenian also belongs to the group of languages, with not yet fully defined set of rules for grapheme to phoneme conversion. It is therefore not possible to assure sufficient accuracy of automatic grapheme to phoneme conversion method. All these difficulties sum up in additional error rate that is introduced in the speech recognition system by the grapheme to phoneme conversion and can be in some cases as high as 30%[1].

One of the possible solutions is to use grapheme based acoustic models, as was proposed by Kanthak and Ney [2]. The grapheme based speech recognisers were also successfully implemented for multilingual and crosslingual speech recognition [3, 4, 5]. The improvements gained by the grapheme based acoustic models mainly depend on the language attributes – the complexity of grapheme to phoneme relations. Current grapheme based acoustic modeling approaches [2, 4] are based on graphemes from the acoustic models' initialization phase on. A new acoustic modeling approach for generation of grapheme based acoustic models is proposed in this paper[1]. The idea is to use the existing phoneme acoustic models, which are converted into the grapheme based form. Thereafter the grapheme based acoustic models can be additionally trained or included directly in the speech recogniser. The first advantage of the proposed method is the possibility of using the existing phoneme acoustic models to start with the training – many research sites have a large sets of high quality phoneme acoustic models. The second advantage of the proposed acoustic modeling method is that are acoustic models generated on wider basis than in the case of standard grapheme based training procedure.

In present ASR systems context dependent acoustic models are usually applied. One of the standard procedures to cope with a large number of free parameters in context dependent acoustic models is the decision tree based clustering [6]. The decision tree is in case of phoneme based acoustic models initialized from phonetic broad classes that are usually manually generated according to acoustic-phonetic properties of phonemes by an expert phonetician. In a last few years a few data-driven methods how to create phonetic broad classes were proposed by different authors [7, 8, 9, 11, 10]. All are having some pros and cons. As acoustic-phonetic properties have no direct representation in the grapheme set, such decision tree initialization is non-representative. In the grapheme acoustic modeling different data-driven methods how to

[1] This work was financially supported by the Slovenian Research Agency under contract number P2-0069.

generate the broad classes were applied [2, 3, 4]. In this paper a data-driven approach for generation of grapheme broad classes based on context-independent grapheme confusion matrix is used. The basic idea for phoneme case was presented in [11]. The procedure is based on a known approach that proves to give good results. It is also well suitable for generating multilingual decision trees, which is of great importancy in present multi-cultural way of communication.

The theoretical background of proposed method how to convert phoneme based acoustic models into grapheme based version is presented in Section 2. The proposed data-driven generation of grapheme broad classes is described in Section 3. More details about the speech database and experimental setup can be found in Section 4 and 5. The evaluation of proposed methods in Section 6 is done as comparison between the acoustic models for phoneme baseline, grapheme baseline using different grapheme broad classes and final grapheme system converted from phoneme version. In this case the decision tree was initialized with the best grapheme broad classes from the grapheme baseline example. The conclusion and directives for future work are given in Section 7.

2. Conversion from phoneme to grapheme acoustic models

The proposed method for generation of grapheme acoustic models builds them from existing phoneme context-independent acoustic models. A grapheme γ is calculated as:

$$\gamma = \sum_{i=1}^{N_p} w_i \phi_i \tag{1}$$

where w_i denotes the weight and the ϕ_i denotes the particular monophone i. The number of monophones included in the conversion is limited with N_p. The Equation (1) can be evolved in such a way that each component of a grapheme acoustic model can be calculated as weighted sum. Thus, the grapheme means $\boldsymbol{\mu}$ are calculated as:

$$\boldsymbol{\mu}_\gamma = \sum_{i=1}^{N_p} w_i \boldsymbol{\mu}_{\phi_i} \tag{2}$$

where $\boldsymbol{\mu}_\gamma$ represents the grapheme mean values, w_i the weight and $\boldsymbol{\mu}_{\phi_i}$ the phoneme mean values. The grapheme variances $\boldsymbol{v}$ are calculated as:

$$\boldsymbol{v}_\gamma = \sum_{i=1}^{N_p} w_i \boldsymbol{v}_{\phi_i} \tag{3}$$

where $\boldsymbol{v}_\gamma$ represents the grapheme variances, w_i the weight and $\boldsymbol{v}_{\phi_i}$ the phoneme variances. The values of transition matrix are calculated as:

$$\alpha_\gamma = \sum_{i=1}^{N_p} w_i \alpha_{\phi_i} \tag{4}$$

where α_γ represents the element of grapheme transition matrix, w_i the weight and α_{ϕ_i} the element of phoneme transition matrix. The last undefined value needed for generation of grapheme acoustic models is the weight w_i. It is defined as:

$$w_i = \frac{con(\gamma, \phi_i)}{\sum_{j=1}^{N_p} con(\gamma, \phi_j)} \tag{5}$$

where $con(\gamma, \phi_i)$ is the number of confusions between the particular phoneme and grapheme. The number of confusions is derived from a phoneme to grapheme confusion matrix, which results from a phoneme speech recogniser, where the development speech set is being transcribed with graphemes.

3. Data-driven generation of grapheme broad classes for decision tree initialization

The applied method for generation of data-driven grapheme broad classes is based on a modified version of the method presented in [11]. The basic idea is that during the speech recognition similar graphemes get confused more often than dissimilar ones. As the decision tree based clustering procedure ties similar states, the grapheme broad classes can be generated on this presumption. The input data can be presented in the form of a grapheme confusion matrix, produced with a context-independent grapheme speech recogniser. The grapheme confusion matrix is generated on some speech data set, which size is approx. one tenth of the full training set. The similarity measure for including a grapheme in the broad class can be defined as:

$$\begin{aligned} \gamma_i \in class_j, \exists con(\gamma_i, \gamma_j) \geq thd_j, \\ 1 \leq i, j \leq G, \end{aligned} \tag{6}$$

where γ_i denotes the current grapheme in the matrix row j. This grapheme is classified into the current class $class_j$. The total number of graphemes is denoted with G, whilst $con(\gamma_i, \gamma_j)$ denotes the number of confusions between current grapheme γ_i and master grapheme γ_j in the matrix row j. The master grapheme in each row j is the one that serves for comparison with all other graphemes in the same row.

The threshold value thd_j in Equation (6) decides, if the grapheme belongs to a particular broad class or not. It is defined as:

$$\begin{aligned} thd_j = \max con(\gamma_i, \gamma_j) weight, \\ 1 \leq i, j \leq G; 0 \leq weight \leq 1, \end{aligned} \tag{7}$$

where thd_j denotes the value of the threshold for the current broad class, $maxcon(\gamma_i, \gamma_j)$ denotes the maximal number of confusions in a matrix row and $weight$ denotes the empirically chosen weight between 0 and 1. In case when the maximal number of confusions is very low (rare grapheme) and low $weight$ is chosen, the threshold thd_j could be 1 or even 0. To prevent such cases, an additional criteria should be defined as:

$$thd_j < I \Rightarrow thd_j = I. \tag{8}$$

When the thd_j value falls below the predefined number I, the thd_j value becomes equal to empirical value I. In such a way, the inclusion of all graphemes in one broad class is hindered.

4. Speech database

All experiments were performed using the Slovenian 1000 FDB SpeechDat(II) fixed telephone database [12]. 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 databases are voice driven telephone applications. The Slovenian SpeechDat(II) database consists from 1000 speaker.

For each speaker 43 different utterances were recorded [13]. Altogether, approximately 30.000 utterances were incorporated in the training set.

Evaluation of generated acoustic models was performed using 6 different test scenarios defined in the SpeechDat(II) database: A1-A6, Q1-Q2, I1, B1C1, O2 and W1-W4 [13]. The test set consisted from 200 speakers. All test scenarios are based on isolated or connected words recognition to exclude the influence of language model on evaluation of acoustic models quality. The original manually corrected phonetic vocabulary provided by the SpeechDat database based on Slovenian SAMPA definition [14] was used for evaluating the phoneme baseline system. The vocabulary for grapheme based cases was generated automatically from orthographic transcription of words. Both vocabularies had a full coverage of test sets.

5. Experimental setup

This paper is focused on different basic acoustic units for speech recognition. To be able to give a fair evaluation of speech recognition performance, the same training procedure should be used for all acoustic model types. In our case, the following procedure, which only differs in the acoustic model type and the version of broad classes, was used.

The feature files produced from speech signal had 12 MFCC coefficient plus the energy. With the first and the second derivative, the final size of 39 elements was achieved. The acoustic models are three state left-right hidden Markov models (HMM), with Gaussian continuous density probability function. The HTK toolkit was used for the experiments [15]. The Slovenian phoneme baseline system had 46 different allophones, while both grapheme based system included 25 Slovenian graphemes.

The first training cycle was based on initialization of all acoustic models with the same global values. Stepwise the number of mixtures was increased to 32 Gaussian. These context-independent acoustic models were used for forced realignment, which results in improved transcriptions of speech material.

The second training cycle was based on particular initialization for each acoustic model. Again, the number of mixtures was stepwise increased. The final context-independent acoustic models, which were used for generating different confusion matrices, had 32 mixtures. The context-dependent acoustic models were refined using the decision tree based clustering procedure, initialized with different types of broad classes. For phoneme acoustic models, the broad classes were defined by an expert, according to acoustic-phonetic properties of Slovenian. For grapheme acoustic models, data-driven broad classes were used. Four different $weight$ values were chosen: $0.05, 0.10, 0.15, 0.20$. The final context-dependent acoustic models used for evaluation had 32 Gaussian mixtures per state and similar complexity.

The conversion from phoneme to grapheme acoustic models was done with the monophone models with 1 Gaussian probability density function per state. The data-driven grapheme broad classes with the best speech recognition performance for grapheme baseline were included in the decision tree based clustering. The most noticeable contribution of phoneme to grapheme conversion method for higher values of N_p was observed for vowel modeling, where particular general vowel (e.g. [a]) has various variants (e.g. long [a:], short [a]).

6. Speech recognition results

Three various types of acoustic models were involved in the speech recognition tests. The phoneme acoustic models were used as baseline. The grapheme acoustic models were first used to determine the best proposed data-driven broad classes, but they were also used as the second baseline. The third type of acoustic models were the proposed grapheme acoustic models converted from existing phoneme acoustic models. All tests were performed with six various test scenarios, described in Section 4. The results are given in the form of the word error rate (WER).

6.1. Phoneme baseline

The first evaluation step was devoted to phoneme baseline (Table 1). The average word error rate $AWER$ is also reported in the last column.

Table 1: *Speech recognition results (WER) for phoneme baseline.*

	A	Q	I	B	O	W	$AWER$
Phn	2.90	0.87	5.70	5.21	7.81	18.56	*6.84*

The achieved results for phoneme baseline are in range of similar ASR systems. The WER was the lowest (0.87%) for the simplest test set with yes/no answers. The most complex test set was W with WER as high as 18.56%. The average word error rate for phoneme baseline system was 6.84%.

6.2. Grapheme baseline with data-driven initialization of decision tree

The grapheme baseline ASR system was also used to evaluate the proposed data-driven method for generation of broad classes. Four different threshold values thd_j were used. The WERs are presented in Table 2.

Table 2: *Word error rates for grapheme baseline with different data-driven broad classes.*

$weight$	A	Q	I	B	O	W	$AWER$
0.05	2.90	0.29	4.15	4.38	6.70	14.97	*5.57*
0.10	2.80	0.29	5.70	4.63	7.25	15.51	*6.03*
0.15	2.90	0.29	4.15	4.56	5.70	15.11	*5.45*
0.20	3.08	0.29	4.15	4.49	6.70	15.24	*5.66*

The performance of grapheme baseline with data-driven broad classes in general improved over the phoneme baseline. The best average WER was 5.45% ($weight = 0.15$), which is a 20.32% relative decrease. The worst grapheme system with $weight = 0.10$ and average WER of 6.03% still performs better (11.84% relative improvement) than phoneme baseline. The analysis of various test set shows similar relation between the scenarios. The best result (0.29% WER) was achieved with the Q test set, while the worst result of 15.51% WER was obtained for the W test set. In some test scenarios different weights produced the best result. This is probably caused by the discrepancy between the development and test set.

6.3. Phoneme to grapheme conversion

The context-dependent converted grapheme acoustic models were trained only with the best data-driven broad class with the threshold value $weight = 0.15$. The conversion from phoneme acoustic models to grapheme acoustic models was done for N_p values 1 and 3. The speech recognition results are presented in Table 3.

Table 3: *Word error rates for converted grapheme acoustic models with N_p values 1 and 3.*

N_p	A	Q	I	B	O	W	$AWER$
1	2.62	0.58	4.15	4.63	5.70	15.16	*5.47*
3	2.15	0.29	4.05	4.38	5.60	14.33	*5.13*

The first row of Table 3 ($N_p = 1$) shows the speech recognition results for the case, where only one phoneme acoustic model was involved in conversion to individual grapheme acoustic model. The achieved performance was very similar to the performance of grapheme baseline with $weight = 0.15$. The best result was again achieved for the Q test set (0.58% WER), while the worst result was with phonetically balanced words - 15.16% WER.

The second row of Table 3 presents the last speech recognition results, with $N_p = 3$. An improvement of performance can be observed. The average word error rate ($AWER$) decreased for 5.87% relatively – from 5.45% to 5.13%. The main improvement was achieved for the A test set (2.15% WER) and W test set (14.33% WER). There was no improvement for the Q test set and B test set.

If we compare the best results with converted grapheme acoustic models with phoneme baseline, it can be seen that the average word error rate decreased for 25.00% relatively. The improvement was most noticeable for the hardest test sets W and O, where the WER decreased from 18.56% to 14.33% and from 7.81% to 5.60%.

Summarization of achieved results shows a clear advantage in the usage of grapheme acoustic models. The improvement in speech recognition performance is even more noticeable in the case, when converted grapheme acoustic models were used. Beside the improvement in performance, such acoustic models have two additional important advantages. The first one is that the phonetic vocabulary is not needed for the testing and the on-line operation. This is very useful for languages, where the grapheme to phoneme conversion presents a non-trivial task. The second advantage is that existing (high quality) phoneme acoustic models can be used for conversion to grapheme acoustic models. In such a way, the training procedure can be simplified and shortened.

The modified method for data-driven generation of grapheme broad classes with confusion matrix also achieved good performance. Already the worst result was better than the phoneme baseline. Data-driven grapheme broad-classes can be very easy adopted to a new grapheme set or even language, as almost no expert knowledge is needed. The disadvantage of the proposed method for data-driven generation of broad classes is the empiric way of selecting the optimal $weight$ value.

7. Conclusion

This paper presented a novel approach for acoustic modeling in speech recognition connected with the usage of grapheme acoustic models. The approach is dealing with conversion from existing phoneme acoustic models to grapheme acoustic models. A significant improvement of speech recognition results was achieved.

The disadvantage of proposed method lies in the nature of grapheme acoustic modeling. Its efficiency is highly tied with the attributes of language and their grapheme to phoneme relationship. The future work will be connected with improvements of conversion procedure and port to multilingual speech recognition environment.

8. References

[1] A. Žgank, and Z. Kačič, ”Comparison of three acoustic basic unit types for Slovenian speech recognition,” *Electrotechnical Review, Journal of Electrical Engineering and Computer Science*, Ljubljana, Slovenia, 2005.

[2] S. Kanthak, and H. Ney, *Context-dependent Acoustic Modeling using Graphemes for Large Vocabulary Speech Recognition.* Proc. ICASSP 2002, Orlando, Florida.

[3] S. Kanthak, and H. Ney, *Multilingual Acoustic Modeling Using Graphemes.* Proc. Eurospeech 2003, Geneva, Switzerland.

[4] M. Killer, S. Stüker, and T. Schultz, *Grapheme Based Speech Recognition.* Proc. Eurospeech 2003, Geneva, Switzerland.

[5] A. Žgank, Z. Kačič, F. Diehl, J. Juhar, S. Lihan, K. Vicsi, and G. Szaszak, *Graphemes as basic units for crosslingual speech recognition.* Proc. ASIDE 2005 Workshop, Aalborg, Denmark.

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

[7] K. Beulen, and H. Ney, *Automatic question generation for decision tree based state tying.* Proc. ICASSP'98, May 1998, pp. 805–808.

[8] R. Singh, B. Raj, and R.M. Stern, *Automatic clustering and generation of contextual questions for tied states in hidden Markov models.*, Proc. ICASSP'99, March 1999, pp. 117–120.

[9] C. Chelba, and R. Morton, *Mutual information phone clustering for decision tree induction.* Proc. ICSLP'2002, Denver, Colorado.

[10] F. Diehl, and A. Moreno, *Acoustic Phonetic Modelling using Local Codebook Features.* Proc. of ICSLP 2004, Jeju Island, Korea.

[11] A. Žgank, B. Horvat, and Z. Kačič, ”Data-driven generation of phonetic broad classes, based on phoneme confusion matrix similarity,” *Speech Communication*, vol. 47, issue 3, pp. 379–393, November 2005.

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

[13] H. van den Heuvel, L. Boves, A. Moreno, M. Omologo, G. Richard, and E. Sanders, ”Annotation in the SpeechDat Projects,” *International Journal of Speech Technology*, vol. 4, issue 2, pp. 127–143, 2001.

[14] http://www.phon.ucl.ac.uk/home/sampa/slovenian.htm

[15] P.C. Woodland, J.J. Odell, V. Valtchev, and S.J. Young, *Large vocabulary continuous speech recognition using HTK.* Proc. ICASSP 1994.