Rule-Based Triphone Mapping for Acoustic Modeling in Automatic Speech Recognition

Sakhia Darjaa[1], Miloš Cerňak[1], Štefan Beňuš[1,2], Milan Rusko[1], Róbert Sabo[1], Marián Trnka[1]

[1]Institute of informatics, Slovak Academy of Sciences
Dúbravská c. 9, 845 07 Bratislava, Slovakia
[2]Department of Eng. and Am. Studies Constantine the Philosopher University, Nitra, Slovakia
{sachia.darzagin,milan.rusko,robert.sabo,milos.cernak,trnka}@savba.sk
sbenus@ukf.sk

Abstract. This paper presents rule-based triphone mapping for acoustic models training in automatic speech recognition. We test if the incorporation of expanded knowledge at the level of parameter tying in acoustic modeling improves the performance of automatic speech recognition in Slovak. We propose a novel technique of knowledge-based triphone tying, which allows the synthesis of unseen triphones. The proposed technique is compared with decision tree-based state tying, and it is shown that for bigger acoustic models, at a size of 3000 states and more, a triphone mapped HMM system achieves better performance than a tree-based state tying system on a large vocabulary continuous speech transription task. Experiments, performed using 350 hours of a Slovak audio database of mixed read and spontaneous speech, are presented. Relative decrease of word error rate was 4.23% for models with 7500 states, and 4.13% at 11500 states.

Key words: automatic speech recognition, acoustic modeling, model tying

1 Introduction

Statistical modeling dominates in current speech technology. In automatic speech recognition (ASR), rare triphones are tied on the model [1] or the state [2] level, and such context modeling based on either data-driven or decision tree clustering significantly improves the recognition performance. It was already shown that the state tying system consistently out-performs the model clustered system.

Phonetic decision tree-based state tying utilizes the knowledge of phonetic classes determining contextually equivalent sets of HMM sets. Facing the challenge of recovering linguistic information in acoustic modeling, which is one of the area for future ASR research specified by [3], we re-visited the process of building the HMM system for Slovak language, showing that our proposed phonetic rule-based triphone tying HMM system outperforms the tree-based state

tying HMM system. The performance gain is achieved with effective triphone mapping, and its latent use in the process of building an HMM system.

The remainder of the paper is structured as follows. In the next Section 2 we introduce rule-based triphone mapping, which we apply to a large-vocabulary continuous speech transcription task in the experimental part of the paper in Section 3. Finally in Section 4 we discuss achieved results.

2 Rule-Based Triphone Mapping

Triphones are context phonemes (basis phoneme P with the left and right context: P_{left}–P+P_{right}). Most triphones are rare and it is not possible to train them robustly. We therefore map rare triphones to more frequent triphones that are much better trained. Thus we constrain contextual information, based on context similarity.

The process of building a triphone mapped HMM system has 4 steps:

1. Training of monophones models with single Gaussian mixtures
2. The number of mixture components in each state is incremented and the models are trained
3. The state output distributions of the monophones are cloned, triphone mapping is applied
4. The triphone tied system is trained again

Unlike the process of building a tied state HMM system [2], monophone models are trained with multiple Gaussian mixtures, and subsequently, state output distributions are cloned for triphone models initialization with a latent application of the triphone map. In a tied state HMM system, cloning and state clustering is done on single Gaussian mixtures, and then the number of mixture components is incremented.

First, the selection of most frequent triphones is performed. Triphones are sorted according to occurrence and a limit is determined. The typical limit from 400 to 800 occurrences is used for databases extending hundred hours. Top N (usually from 2000 to 3500), most frequent triphones, are thus selected from all available contexts. As mapping is not applied to context-free phonemes, such as *sp* and *sil*, they are added to the selection list as monophones. If there are less frequent phonemes that are not represented in the middle part of triphones, these are added to the selection list as well.

2.1 Rules for phonetic similarity in Slovak

We used a discrete rule-based approach for determining an ordered list of candidate phonemes for each of the 45 target phonemes of Slovak. These candidate lists are ordered based on the phonetic distance from the target phoneme to the candidate phoneme. The ordering process was based on several basic principles that are motivated by general and Slovak-specific phonetic considerations and on strategies for resolutions if the principles are in conflict or if they are not

sufficient for uniquely populating the ordered lists. Fig. 1 enlists the ordered 10 (for illustration) candidate phonemes for a set of selected target phonemes.

		Candidates									
	Target	1	2	3	4	5	6	7	8	9	10
Syllabic	a	a	o	r	ɛ	l	u	ĭa	i	aː	ĭe
	aː	aː	a	o	ɛ	oː	ɛː	uː	r	ĭa	ĭe
	ɛ	ɛ	i	iː	ĭe	a	o	j	ĭa	ĭu	ŏo
	i	i	ɛ	j	n	m	r	u	iː	uː	ĭe
	o	o	a	u	ɛ	ŏo	r	ĭe	ĭu	ĭa	l
	u	u	o	i	oː	ĭu	r	ɛ	l	iː	a
	ĭe	ĭe	ɛ	ĭa	ĭu	ŏo	a	o	u	l	i
	l	l	l	r	rː	o	a	ɛ	i	u	ĭa
	r	r	r	l	rː	a	o	ɛ	i	u	aː
Consonantal	b	b	g	d	p	r	j	m	n	l	ɟ
	d	d	ɟ	b	g	m	n	dz	z	ɦ	r
	dʒ	dʒ	dz	d	ɟ	ɦ	c	s	k	t	ts
	f	f	t	p	k	x	dz	z	s	ɦ	ts
	g	g	b	d	dz	ɟ	n	m	r	j	ɦ
	k	k	t	p	x	ɦ	g	f	d	c	ɟ
	n	n	ɲ	m	ɟ	b	d	dz	g	j	r
	p	p	t	k	c	f	b	x	r	v	ɟ
	s	s	ʃ	f	ts	tʃ	x	t	k	p	c
	ʃ	ʃ	tʃ	s	f	ts	dʒ	dz	x	ʒ	p
	t	t	p	k	c	f	ts	dz	ɟ	s	tʃ
	tʃ	tʃ	ts	f	s	ʃ	z	ʒ	dʒ	k	p

Fig. 1. Ordered canditate list for selected target phonemes.

The principles are violable (in the sense of constraints of Optimality Theory [4], and range from general, such as the preservation of the identity of segment types (consonant, vowel) or the possibility of substitution between vowels and sonorants, to more specific ones, such as the preference for the identity of the manner of articulation and voicing in consonants over the identity in the place of articulation, the preference of preserving vowel height over its frontness, or the preference for the substitution of target diphthongs with those candidate monophthongs that are identical to the second element of diphthongs. This last strategy is motivated by the fact that Slovak has so called rising diphthongs in which the 2^{nd} element is more prominent than the first one.

The heuristic strategies, which filled the gaps or resolved the conflicts after the application of the principles, were based on minimizing the effect of a substitution on the surrounding phonemes taking into account both acoustic and articulatory considerations. For example, /r/ is acoustically the most similar to a schwa-like vowel, thus affects format transitions minimally, and articulatorily involves only a brief tongue tip gesture, thus minimally affecting the tongue body as the main vocalic articulator. Both of these features play an important role in a relatively close proximity (i.e. low rank) of /r/ in the candidate lists for the back vowels /a/ and /i/.

```
for each triphone Pi-P+Pj (incl. unseen) do:
  for each triphone Pm-P+Pn from the selected triphones with
  the same basis phoneme P do:
    perform left context mapping:
      target_phoneme = Pi
      candidate_phoneme = Pm
      left_context = position of candidate_phoneme in the list
                     belonging to target_phoneme
    perform right context mapping:
      target_phoneme = Pj
      candidate_phoneme = Pn
      right_context = position of candidate_phoneme in the list
                      belonging to target_phoneme
    context_tying_cost = left_context + right_context
  perform triphone mapping:
    Pi-P+Pj is mapped to Pm-P+Pn with minimal context_tying_cost
    if there are more Pm-P+Pn with minimal context_tying_cost do:
      for each Pm-P+Pn do:
        if left_context < right_context do:
          Pi-P+Pj is mapped to Pm-P+Pn
```

Fig. 2. Algorithm of triphone mapping. Having a single mapped triphone from the list of all triphones, P_i–P+P_j, and multiple mapping candidate triphones P_m–P+P_n with the same basis phoneme P, the candidate triphone with minimal context tying cost and better left context mapping is selected.

The resulting discrete matrix of partial phoneme confusions thus provides an input for the algorithm that uses the phoneme distance (a position of candidate phoneme in target phoneme row) and subsequently maps each triphone into the closest triphone from the list of selected triphones with the same basis phoneme. The task of triphone mapping consists of separate left context and right context mapping, based on the basic premise that the left context is more important then the right context. Fig. 2 presents the algorithm of context tying for triphone mapping in meta programing language.

3 Experiments

The aim of the experiment was to compare tree-based state tying with triphone mapping systems (data-driven state clustering [5] was not considered, as it does not allow synthesis of unseen triphones). Both systems were trained using the same number of Baum-Welch re-estimations, the same number of Gaussian mixtures, and used the same initial set of untied triphones. The tree-based state tying system was trained according to [2] and [6] training procedures. The triphone mapped system was then created according to Sec. 2.

Julius decoder [7] was used as a reference speech recognition engine, and the HTK toolkit was used for word-internal acoustic models training. A set of phonetic questions used in decision trees was taken from the multi-lingual system [8], where the Slovak system achieved state-of-the-art performance when compared to other participating languages. To gain some impression of used questions, Tab. 1 shows the criteria for phonetic grouping used in decision trees.

Table 1. *The criteria for phonetic grouping used for questions in tree-based state tying in Slovak speech recognition system. Both right (R) and left (L) contexts were considered.*

Vowels	Consonants
R,L-short	R,L-sonants
R,L-long	R,L-plosives; voiced/unvoiced
R,L-monophtongs	R,L-fricatives; voiced/unvoiced
R,L-diphtongs	R,L-affricatives; voiced/unvoiced
R,L-front	R,L-labial
R,L-back	R,L-glottal
R,L-open, closed	R,L-lingual
R,L-halfopen	R,L-unvoiced

3.1 Data

Experiments have been performed using both read and spontaneous speech databases of the Slovak language. The first database contained 250 hours of gen-

der balanced read speech, recorded from 250 speakers with a Sennheiser ME3 Headset Microphone with an In-Line Preamplifier Sennheiser MZA 900 P. The second database contained 100 hours of 90% male spontaneous speech, recorded from 120 speakers at council hall with goose neck microphones. Databases were annotated using the Transcriber annotation tool [9], twice checked and corrected. Whenever possible, recordings were split into segments not bigger than 10 sec. Our testing corpus contained 20 hours of recordings obtained by randomly selecting segments from each speaker contained in the first read speech database. These segments were not used in training.

A text corpus was created using a system that retrieves text data from various Internet pages and electronic sources that are written in the Slovak language. Text data were normalized by additional modifications such as word tokenization, sentence segmentation, deletion of punctuation, abbreviation expanding, numerals transcription, etc. The system for text gathering also included constraints such as filtering of grammatically incorrect words by spellchecking, duplicity verification of text documents and others constraints. The text corpora contained a total of about 92 million sentences with 1.25 billion Slovak words. Trigram language models (LMs) were created with a vocabulary size of 350k unique words (400k pronunciation variants) which passed the spellcheck lexicon and subsequently were also checked manually. As a smoothing technique the modified Kneser-Ney algorithm was used [10].

3.2 Results

First, we trained acoustic models (AMs) using tree-based state tying. By setting 1) the outlier threshold that determines the minimum occupancy of any cluster (RO command), and 2) the threshold of the minimal increase in log likelihood achievable by any question at any node of the decision tree (the first argument of TB command), we trained four AMs with different numbers of states (in the range from 2447 to 11489).

Next, we trained AMs using the proposed triphone mapping as described in Sec. 2. In order to achieve the same range of trained states, we set the limit of minimum occupancy N of triphones in the range from 200 to 2850.

Fig. 3 shows the results of tree-based state tying and triphone mapping. For small acoustic modeling up to 3000 states, tree-based state tying slightly outperforms triphone mapping (e.g. for models with 2500 states). For bigger models, at the size typical for large vocabulary continuous speech recognition (LVCSR) training (more than 3000 states), rule-based triphone mapping achieves better word error rate (WER). Relative decrease of word error rate was 4.23% for models with 7500 states, and 4.13% with 11500 states.

4 Discussion

We showed that the rule-based triphone mapped HMM system achieves better WER for models typical for LVCSR training. This result poses an interesting

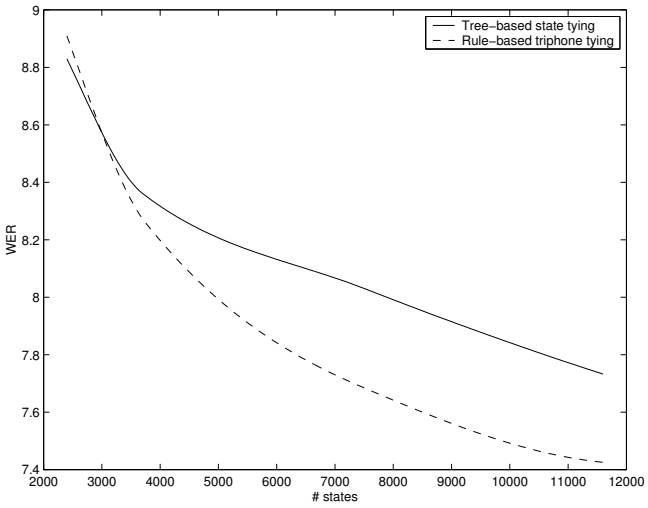

Fig. 3. Tree-based state tying compared to triphone tying on the number of trained HMM states. Four acoustic models were trained for each function with 2450, 3700, 7450 and 12000 states. The results were then interpolated to get smooth functions.

question: Why does the triphone mapped HMM system, the model tying approach, perform better than the state tying approach, if it was already shown that state tying systems consistently out-perform model clustered systems (see e.g. [2])?

In the process of building the triphone mapped HMM system we emphasized, that while the standard state tying system clusters the states from the single Gaussian models (due to performance reasons), trained roughly in 1/3 from all the training time, the triphone mapped system can cluster the models from the multiple Gaussian Mixture Models (GMMs), trained roughly in 4/5 from all the training time. Both tying systems work with single Gaussian models for the calculation of distance metric; however, the triphone mapping can be easily applied later in the training process, when monophone models are much better trained using multiple Gaussians. In order to verify this hypothesis, we forced the process of building the triphone mapped HMM system to be more similar to the building the state tying system (cloning and clustering the triphones from the single Gaussian models):

1. Training of monophones models with single Gaussian mixtures
2. The state output distributions of the monophones are cloned, triphone mapping is applied
3. The triphone tied system is trained
4. The number of mixture components in each state is incremented and the models are trained again

We trained the triphone mapped HMM system using this modified process above, and for 12000 states we obtained similar performance as with the state tying system and the same model size. We can thus conclude that the main gain in performance is due to latent application of triphone mapping. Having well trained monophones using multiple Gaussians distributions, the cloned triphones are better initialized than with single Gaussians monophones. The performance change at 3000 states is probably related to the amount of training data available. The more data we have, the more states we can robustly train.

The process of triphone mapping is language independent, and can be further tuned with an application of different weights for the left and right contexts. Data-driven triphone mapping belongs to our future work as well.

Acknowledgements This work was supported in part by the EU grant (ITMS 26240220064).

References

1. Bahl, L.R., deSouza, P.V., Gopalakrishnan, P.S., Nahamoo, D., Picheny, M.A.: Decision trees for phonological rules in continuous speech. In: ICASSP-91. ICASSP '91, Washington, DC, USA, IEEE Computer Society (1991) 185–188
2. Young, S.J., Odell, J.J., Woodland, P.C.: Tree-based state tying for high accuracy acoustic modelling. In: Proceedings of the workshop on Human Language Technology. HLT '94, Stroudsburg, PA, USA, ACL (1994) 307–312
3. Baker, J., Deng, L., Khudanpur, S., Lee, C.H., Glass, J., Morgan, N., O'Shaughnessy, D.: Updated MINDS report on speech recognition and understanding, Part 2 [DSP Education]. IEEE Signal Processing Magazine **26**(4) (July 2009) 78–85
4. Prince, A., Smolensky, P.: Optimality Theory: Constraint Interaction in Generative Grammar. Blackwell (1993/2004)
5. Young, S., Woodland, P.C.: State clustering in hidden Markov model-based continuous speech recognition. Computer Speech & Language **8**(4) (October 1994) 369–383
6. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Ovell, J., Ollason, D., Valtchev, D.P.V., Woodland, P.: The HTK Book (for v3.4.1). Cambridge (2009)
7. Lee, A., Kawahara, T., Shikano, K.: Julius – an Open Source Real-Time Large Vocabulary Recognition Engine. In: Proc. of the European Conference on Speech Communications and Technology (EUROSPEECH), Aalborg, Denmark (September 2001)
8. Johansen, F.T., Warakagoda, N., Lindberg, B., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., Salvi, G.: The COST 249 SpeechDat multilingual reference recogniser. In: Proc. of the 2nd Intl. Conf. on LREC, Athens (May 2000)
9. Barras, C., Geoffrois, E., Wu, Z., Liberman, M.: Transcriber: development and use of a tool for assisting speech corpora production. Speech Communication **33**(1–2) (January 2000)
10. Staš, J., Hládek, D., Juhár, J.: Language Model Adaptation for Slovak LVCSR. In: Proc. of the Intl. Conference on AEI, Venice, Italy (2010) 101–106