Effective Triphone Mapping for Acoustic Modeling in Speech Recognition

Sakhia Darjaa, Miloš Cerňak, Marián Trnka, Milan Rusko, Róbert Sabo

Institute of informatics, Slovak Academy of Sciences
Dúbravská c. 9, 845 07 Bratislava
{darjaa,milos.cernak,milan.rusko,trnka,robert.sabo}@savba.sk

Abstract

This paper presents effective triphone mapping for acoustic models training in automatic speech recognition, which allows the synthesis of unseen triphones. The description of this data-driven model clustering, including experiments performed using 350 hours of a Slovak audio database of mixed read and spontaneous speech, are presented. The proposed technique is compared with tree-based state tying, and it is shown that for bigger acoustic models, at a size of 4000 states and more, a triphone mapped HMM system achieves better performance than a tree-based state tying system. The main gain in performance is due to latent application of triphone mapping on monophones with multiple Gaussian pdfs, so the cloned triphones are initialized better than with single Gaussians monophones. Absolute decrease of word error rate was 0.46% (5.73% relatively) for models with 7500 states, and decreased to 0.4% (5.17% relatively) gain at 11500 states.

Index Terms: automatic speech recognition, acoustic modeling, model tying

1. Introduction

Eastern European languages spoken by smaller populations [1], such as the Slovak language, can be considered as under–resourced, as they suffer from the lack of audio databases and linguistic resources. Focused attempts in speech recognition in Slovak have begun at the Department of speech synthesis and analysis of the Slovak Academy of Sciences with SpeechDat-E telephone speech database creation 10 years ago (see for example [2]). Robust model training that copes with limited data belongs to one of our biggest challenges.

Statistical modeling dominates in current speech technology. In automatic speech recognition (ASR), rare triphones are tied on model [3] or state [4] 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-perform the model clustered system.

We re-visited the process of building the HMM system for Slovak language, proving that the tree-based state tying system does not achieve the same accuracy level as our novel model tying system. The performance gain is achieved with effective triphone mapping, and its latent use in the process of building an HMM system – standard tree-based state tying technique clusters states of single Gaussian monophones, while we propose cloning Gaussian Mixture Models (GMMs) state output distributions of the monophones for all the triphones and then apply triphone clustering. The proposed triphone mapped system out-performs tree-based state tying for acoustic models of 4k states and more. For smaller models with less than 4k states the performance is equal.

The paper is organized as follows: Section 2 proposes triphone mapping using context similarity and its using in HMM training. Section 3 describes performed experiments and Section 4 discuss the results.

2. Triphone Mapped HMM System

Triphones are context phonemes (basis phoneme P with the left and right context: P_{left}–P+P_{right}). Most of 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 multiple GMMs are trained.
3. The state output distributions of the monophones are cloned for all triphones, triphone mapping for clustering is applied
4. The triphone tied system is trained.

Unlike the process of building a tied state HMM system [4], 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 due to performance considerations done on single Gaussian mixtures, and then the number of mixture components is incremented.

2.1. Triphone mapping

In this section we describe data-driven model clustering for triphone mapping. Unlike data-driven method of state clustering [5], proposed triphone mapping allows the synthesis of unseen triphones.

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.

We used a distance of HMM states of individual 3-state monophones for mapping of all triphones to the selected triphones. The distance of a phoneme to itself is 0 for all states. The acoustic distance AD is calculated from single mixture Gaussians monophones for each phone pair and HMM state - 1 (left), 2 (middle) and 3 (right). The distance of two phonemes i and j is calculated as:

$$AD_{HTK}(i,j) = \sqrt{\frac{1}{V_f} * \sum_{k=1}^{V_f} \frac{(\mu_{ik} - \mu_{jk})^2}{\sigma_{ik}\sigma_{jk}}} \quad (1)$$

where V_f is the dimensionality of feature vector f, μ and σ are means and variances, respectively. The distance is calculated for each emmiting state, resulting in i–j–1, i–j–2 and i–j–3 values for the phoneme pair. The Eq. 1 is a simplified distance metric of the square root of the divergence between the two Gaussian pdfs [5], and it is the Eq. (17.1) from HTKbook [6] for a single data stream.

Alternatively, inspired by other works such as [7], we investigated Bhattacharyya distance as well:

$$\begin{aligned} AD_{Bhat}(i,j) = & \sum_{k=1}^{V_f} \frac{1}{4} \frac{(\mu_{ik} - \mu_{jk})^2}{\sigma_{ik} - \sigma_{jk}} \\ & + \frac{1}{2} \log \left| \frac{\sigma_{ik}+\sigma_{jk}}{2} \right| \\ & - \frac{1}{4} \log |\sigma_{ik}\sigma_{jk}| \end{aligned} \quad (2)$$

The triphone distance TD between two triphones is defined by following Eq. (3) and (5). The triphone distance TD_2 uses left and right contexts:

$$\begin{aligned} TD_2(P_i\text{–}P\text{+}P_j, P_k\text{–}P\text{+}P_l) = \\ w(P, \text{left}) AD(P_i\text{–}P_k\text{–}3) \\ + w(P, \text{right}) AD(P_j\text{–}P_l\text{–}1) \end{aligned} \quad (3)$$

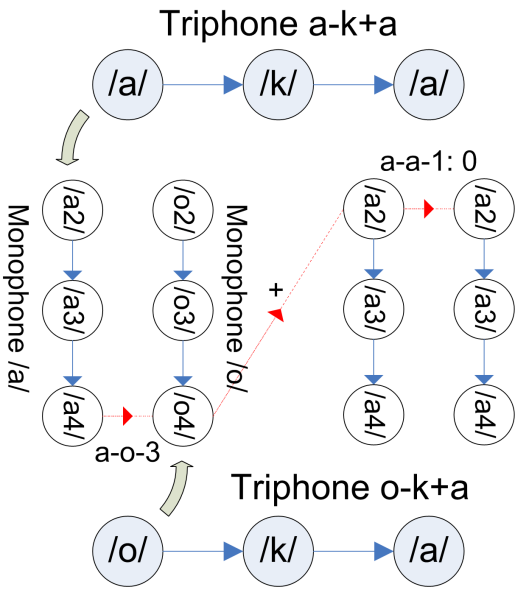

Figure 1: Triphone distance TD_2 calculation of /a-k+a/ and /o-k+a/, assuming the same basis phoneme /k/. The dotted line represents addition of AD of a-o-3 and the second addition of a-a-1 distance (zero - the distance of /a/ to itself). Monophone states, single-Gaussian models calculated in Step 1, are white. Triphone states, GMMs calculated in Step 2, are darker.

where $w(.)$ are context weights, and AD is acoustic distance defined by Eq. (1) – (2). Fig. 1 shows the calculation of the triphone distance graphically. We then map each triphone m from the list of all possible triphones, which includes unseen triphones as well, into the closest triphone n^* from the list of selected triphones with the same basis phoneme:

$$n^* = TriMap(m) = \underset{n}{\text{argmin}}(TD_2(m,n)), \quad (4)$$

where $n \in \{n_1, \ldots, n_N\}$. A generalized distance TD_3 might be calculated from all three parts of the triphones. The distance with different basis phonemes P_m and P_n is then defined as:

$$\begin{aligned} TD_3(P_i\text{–}P_m\text{+}P_j, P_k\text{–}P_n\text{+}P_l) = \\ w(\text{left}) AD(P_i\text{–}P_k\text{–}3) \\ + w(\text{mid}) AD(P_m\text{–}P_n\text{–}2) \\ + w(\text{right}) AD(P_j\text{–}P_l\text{–}1) \end{aligned} \quad (5)$$

The weights are thus not tied to a concrete basis phoneme, as $P_m \neq P_n$.

3. Experiments

The aim of the experiment was to compare standard 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 [4] and [6] training recipes. The triphone mapped (using Eq. (4)) system was then created according to Sec. 2.

Julius decoder [8] 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 multilingual system [9], 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 Slovak language. The first database contained 250 hours of gender balanced read speech, recorded from 250 speakers with Sennheiser ME3 Headset Microphone with 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 Transcriber annotation tool [10], twice checked and corrected. Recordings were split into segments if possible not bigger than 10 sec. Testing corpus contained 20 hours of recordings obtained by random selection segments from each speaker from 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 [11].

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. 2 shows the results of tree-based state tying and triphone mapping for TD_2 triphone distance and two distances AD_{HTK} and AD_{Bhat}. The context weights $w(P, \text{left})$ and $w(P, \text{right})$ were set to 1. For acoustic modeling up to 4000 states, tree-based state tying is equal or slightly outperforms triphone mapping (e.g. for models with 3700 states). For bigger models, at the size typical for large vocabulary continuous speech recognition (LVCSR) training (more than 4000 states), triphone mapping achieves better word error rate (WER). Absolute decrease of WER using D_{HTK} mapping was 0.46% (5.73% relatively) for models with 7500 states, and decreased to 0.4% (5.17% relatively) gain with 11500 states.

4. Discussion

We showed that the triphone mapped HMM system achieves better accuracy for models typical for LVCSR training. This result poses an interesting question: Why does the triphone mapped HMM system, the model tying approach, perform better than state tying approach, if it was already shown that the state tying consistently out-perform the model clustered system (see e.g. [4])?

The state tying system is due to performance considerations forced to cluster the states from the single Gaussian models (there is no closed form for calculating the KL Divergence between GMMs but there are many

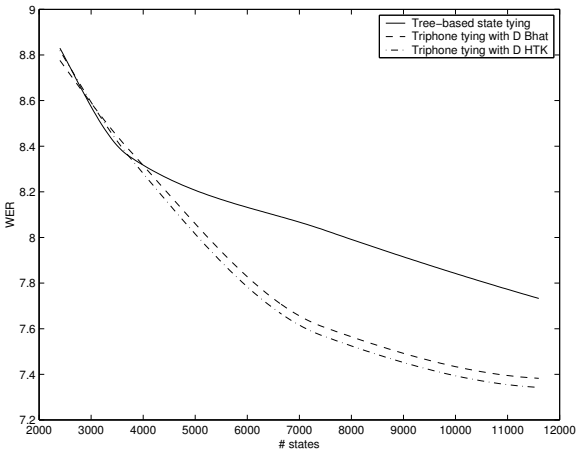

Figure 2: Tree-based state tying compared to triphone tying on number of trained HMM states. Four acoustic models were trained for each function with 2450, 3700, 7450 and 12000 states. The results were then interpolated to get smooth functions.

approximations, see e.g. [12]), trained roughly during $1/3$ of all the training time, the triphone mapped system can easily cluster the models from the multiple GMMs, trained roughly during $4/5$ of all the training time. Both tying systems thus work with single Gaussian models for the calculation of distance metric; however, the triphone mapping is 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 of the state tying system (cloning and clustering the triphones from the single Gaussian models):

1. Training of monophones models with single Gaussians.
2. The state output distributions of the monophones are cloned for all triphones, triphone mapping for clustering 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 got WER of 7.73%. The state tying system for this model had WER of 7.74%. 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 4000 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 contextual weights $w(.)$ in Eq. (3), (5), and a verification of generalized triphone distance TD_3. Moreover, we work on triphone mapping fully based on expert rules, as the triphone mapping technique allows to recover linguistic information in acoustic modeling, which is one of the area for future ASR research specified by [13].

5. Acknowledgements

We would like to thank Richard Kováč, who helped us with initial versions of the triphone mapping.

The work has been partially funded by the EU grant (ITMS 26240220060).

6. References

[1] S. Krauwer, "The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap," in *Proc. of SPECOM*, Moscow, Russia, Sept. 2003.

[2] M. Rusko, S. Daržagín, and M. Trnka, "SpeechDat-E telephone speech database as an important source for basic acoustic-phonetic research in Slovak," in *Proc. of the Intl. Congress on Acoustics (ICA 2004)*, Kyoto, Japan, 2004, pp. II–1676–II–1682.

[3] L. R. Bahl, P. V. deSouza, P. S. Gopalakrishnan, D. Nahamoo, and M. A. Picheny, "Decision trees for phonological rules in continuous speech," in *ICASSP-91*, Washington, DC, USA, 1991, ICASSP '91, pp. 185–188, IEEE Computer Society.

[4] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology*, Stroudsburg, PA, USA, 1994, HLT '94, pp. 307–312, ACL.

[5] S. Young and P. C. Woodland, "State clustering in hidden Markov model-based continuous speech recognition," *Computer Speech & Language*, vol. 8, no. 4, pp. 369–383, Oct. 1994.

[6] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Ovell, D. Ollason, D. Povey V. Valtchev, and P. Woodland, *The HTK Book (for v3.4.1)*, Cambridge, 2009.

[7] B. Mak and E. Barnard, "Phone clustering using the Bhattacharyya distance," in *ICSLP*, October 1996, pp. 2005–2008.

[8] A. Lee, T. Kawahara, and K. Shikano, "Julius – an Open Source Real-Time Large Vocabulary Recognition Engine," in *Proc. of the European Conference on Speech Communications and Technology (EUROSPEECH)*, Aalborg, Denmark, Sept. 2001.

[9] F. T. Johansen, N. Warakagoda, B. Lindberg, G. Lehtinen, Z. Kačič, A. Žgank, K. Elenius, and G. Salvi, "The COST 249 SpeechDat multilingual reference recogniser," in *Proc. of the 2nd Intl. Conf. on LREC*, Athens, May 2000.

[10] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication*, vol. 33, no. 1–2, Jan. 2000.

[11] J. Staš, D. Hládek, and J. Juhár, "Language Model Adaptation for Slovak LVCSR," in *Proc. of the Intl. Conference on AEI*, Venice, Italy, 2010, pp. 101–106.

[12] John R. Hershey and Peder A. Olsen, "Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models," in *ICASSP*, 2007, vol. 4, pp. 317–320.

[13] J. Baker, Li Deng, S. Khudanpur, Chin-Hui Lee, J. Glass, N. Morgan, and D. O'Shaughnessy, "Updated MINDS report on speech recognition and understanding, Part 2 [DSP Education]," *IEEE Signal Processing Magazine*, vol. 26, no. 4, pp. 78–85, July 2009.