Initialization of Adaptation by Sufficient Statistics Using Phonetic Tree

Zbyněk Zajíc, Lukáš Machlica, Luděk Müller
Department of Cybernetics, Faculty of Applied Sciences
University of West Bohemia, Plzeň, Czech Republic
Email: {zzajic, machlica, muller}@kky.zcu.cz

***Abstract*—In this work we deal with the problem of small amount of data when estimating a feature transformation for the speaker adaptation of an acoustic model. Our goal is to compensate for the lack of adaptation data by a proper initialization of transformation matrices. Methods used in such situations are described, they are based on collecting additional accumulated statistics from nearest speakers. The proposed initialization approach is based on accumulated statistics too, but it incorporates also phonetic information when selecting the "nearest" statistics. Initialization methods compensating for the absence of actual speaker's data are tested on telephone recordings with different amounts of adaptation data. In worst situation with extremely small amount of adaptation data relative improvement of 5% is obtained.**

Keywords: speech recognition, adaptation, initialization, phonetic tree

I. Introduction

Nowadays, a speaker adaptation of an acoustic model is a standard approach used to improve the accuracy of Automatic Speech Recognition (ASR) systems. In this work focus will be given on feature Maximum Likelihood Linear Regression (fMLLR) adaptation (1). fMLLR, like other adaptation techniques, utilizes only some statistics of adaptation data referred in Section II. In cases where small amount of adaptation data is available the number of free parameters to be estimated becomes easily too high to estimate these parameters reliably. Thus, fMLLR transformation matrix becomes ill-conditioned, what can lead to poor recognition rates. Various solutions to avoid this problem have been proposed, e.g. lowering the number of free parameters by using diagonal or block diagonal transformation matrices (2) or finding transformation matrices as a linear combination of basis matrices (3). Another solution is performing a proper initialization of transformation matrices (4), (5).

Initialization methods compensate for the absence of actual speaker's data, however the initialization suppresses the influence of actual adaptation data for the benefit of initialization data. Usually a compromise has to be made between stability and accuracy of the adaptation. The problem how to choose suitable initialization data is addressed in Section III. One of the solutions is to use initialization derived directly from a SI model (4). A better alternative is to use data from the development set that are acoustically close to the actual speaker. This method was introduced in (5), it collects additional statistics from "nearest" speakers from the development set, and subsequently these statistics are added to the statistics from the actual speaker.

In this work we will go further, we will distinguish not only similarities between speakers but also the dissimilarities between phonetic units of this speaker. Thus, speech of each speaker will be divided to phonetic classes, for each class of each speaker initialization statistics will be accumulated, and the "nearest" statistics for initialization will be composed only of parts of the speech from development speakers (not of the whole speech as was done before), details are given in Section III-B). Such statistics proved to be more suitable to compensate for the absence of adaptation data. Results for different initialization types and classic (basic/uninitialized) fMLLR for different amounts of adaptation data can be found in Section IV. Contribution of proposed methods was demonstrated on spontaneous telephone speech.

II. Adaptation

The difference between adaptation and ordinary training methods stands in the prior knowledge about the distribution of model parameters, usually derived from the SI model. The adaptation adjusts the SI model so that the probability of the adaptation data would be maximized. Adaptation techniques do not access the data directly, but only through accumulated statistics, which is the first step preceding the adaptation process.

A. Adaptation Statistics

Instead of storing a huge amount of data adaptation methods need only following statistics:

$$\gamma_{jm}(t) = \frac{\omega_{jm} p(\boldsymbol{o}_t|jm)}{\sum_{m=1}^{M} \omega_{jm} p(\boldsymbol{o}_t|jm)} \tag{1}$$

standing for the m-th mixtures' posterior of the j-th state of the HMM for feature vector $\boldsymbol{o}_t$,

$$c_{jm} = \sum_{t=1}^{T} \gamma_{jm}(t) \tag{2}$$

978-1-4673-2197-6/12/$31.00 ©2012 IEEE

representing the soft count of mixture m,

$$\varepsilon_{jm} = \sum_{t=1}^{T} \gamma_{jm}(t)\boldsymbol{o}_t,\ \varepsilon^2_{jm} = \sum_{t=1}^{T} \gamma_{jm}(t)\boldsymbol{o}_t\boldsymbol{o}_t^{\mathrm{T}}, \tag{3}$$

denoting the sum of the first and the second moment of features aligned to mixture m in the j-th state of the HMM.

B. Feature Maximum Likelihood Linear Regression

fMLLR try to find a linear transformation in order to match adaptation data with an acoustics model. This technique can adapt more model components at once using the same transformation (e.g. only one matrix for all the model means). Similar model components are clustered into clusters $K_n, n = 1,\ldots,N$ in order to lower the number of adapted parameters (6). fMLLR transforms features $\boldsymbol{o}_t$ according to

$$\bar{\boldsymbol{o}}_t = \boldsymbol{A}_{(n)}\boldsymbol{o}_t + \boldsymbol{b}_{(n)} = \boldsymbol{W}_{(n)}\boldsymbol{\xi}(t)\,, \tag{4}$$

where

$$\boldsymbol{W}_{(n)} = [\boldsymbol{A}_{(n)}, \boldsymbol{b}_{(n)}], \tag{5}$$

$\boldsymbol{W}_{(n)}$ represents the transformation matrix corresponding to the n-th cluster K_n and $\boldsymbol{\xi}(t) = [\boldsymbol{o}_t^{\mathrm{T}}, 1]^{\mathrm{T}}$ stands for the extended feature vector.

The estimation formulas of rows of $\boldsymbol{W}_{(n)}$ are given as

$$\boldsymbol{w}_{(n)i} = \boldsymbol{G}^{-1}_{(n)i}\left(\frac{\boldsymbol{v}_{(n)i}}{\alpha_{(n)}} + \boldsymbol{k}_{(n)i}\right), \tag{6}$$

where $\boldsymbol{v}_{(n)i}$ is the i-th row vector of cofactors of matrix $\boldsymbol{A}_{(n)}$, $\alpha_{(n)}$ can be found as a solution of a quadratic function

$$\beta_{(n)}\alpha^2_{(n)} - \alpha_{(n)}\,\boldsymbol{v}^{\mathrm{T}}_{(n)i}\boldsymbol{G}^{-1}_{(n)i}\boldsymbol{k}_{(n)i} - \boldsymbol{v}^{\mathrm{T}}_{(n)i}\boldsymbol{G}^{-1}_{(n)i}\boldsymbol{v}_{(n)i} = 0 \tag{7}$$

and

$$\boldsymbol{k}_{(n)i} = \sum_{m\in K_n} \frac{\mu_{mi}\varepsilon_m(\boldsymbol{\xi})}{\sigma^2_{mi}}, \tag{8}$$

$$\boldsymbol{G}_{(n)i} = \sum_{m\in K_n} \frac{\varepsilon_m(\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}})}{\sigma^2_{mi}}, \tag{9}$$

$\boldsymbol{G}_{(n)i}$, $\boldsymbol{k}_{(n)i}$ are accumulation matrices of statistics (3) of all mixtures m contained in a given cluster K_n, and

$$\varepsilon_m(\boldsymbol{\xi}) = \left[\varepsilon_m^{\mathrm{T}}, c_m\right]^{\mathrm{T}}, \tag{10}$$

$$\varepsilon_m(\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}) = \begin{bmatrix} \varepsilon^2_m & \varepsilon_m \\ \varepsilon_m^{\mathrm{T}} & c_m \end{bmatrix}. \tag{11}$$

Equation (6) is a solution of the minimization problem with auxiliary function given in (7).

Note that the new estimate of $\boldsymbol{w}_{(n)i}$ given in (6) depends on the previous estimate of $\boldsymbol{W}_{(n)}$ through $\boldsymbol{v}_{(n)i}$ and $\alpha_{(n)}$. Thus, several iterations have to be run to acquire convergence of parameters of the matrix $\boldsymbol{W}_{(n)} = [\boldsymbol{A}_{(n)}, \boldsymbol{b}_{(n)}]$. The starting matrices $\boldsymbol{A}_{(n)}$, $\boldsymbol{b}_{(n)}$ of the iteration process have to be chosen (e.g. random matrices).

III. Initialization

The matrices of accumulated statistics $\boldsymbol{G}_{(n)i}$, $\boldsymbol{k}_{(n)i}$ are dense and if only low amount of data is available they can lead to ill-conditioned transformation matrices $\boldsymbol{W}_n$. To avoid the degradation of system's performance matrices $\boldsymbol{G}_{(n)i}$, $\boldsymbol{k}_{(n)i}$ have to be suitably initialized. One of the possibilities proposed in (4) is restrict the influence of the adaptation data and reduce the influence of the adaptation – resulting model is closer to the SI model (when none new adaptation data are available, the estimated transformation matrix should equal the identity matrix). This method utilizes directly the SI model parameters.

Rather than to restrict the influence of adaptation we replace the absence of test data with statistics collected from similar speakers. Note that the term *test* data (speaker) denotes the data (speaker), which are to be adapted.

A. Sufficient Statistics from N-best Speakers

To collect statistics from speakers similar to the given test speaker a development set containing a lot of speakers with different voices is used, and their statistics are utilized for initialization (5), see Fig.1.

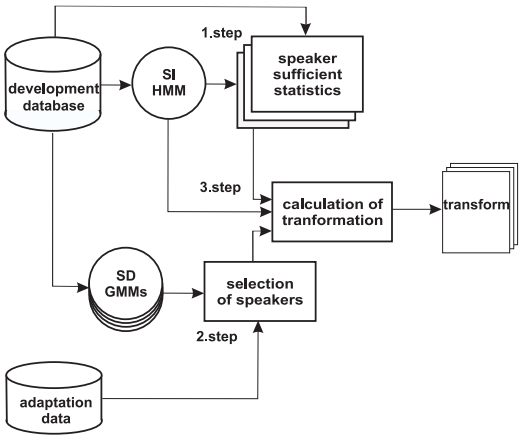

Fig. 1. Three steps in the process of fMLLR model adaptation based on additional statistics. SI stands for Speaker Independent model, SD for Speaker Dependent model.

The algorithm can be divided to three steps:

- **Accumulation of statistics** – for each speaker s from the development set matrices $\boldsymbol{k}^s_{(n)i}$ and $\boldsymbol{G}^s_{(n)i}$ given in (8), (9) are accumulated and stored off-line. Also a Gaussian Mixture Model (GMM) is trained for each development speaker.
- **Selecting a cohort of speakers** – N-best speakers are selected from the development database according to their closeness to the test speaker. The "nearest" speakers are determined according to methods of speaker recognition (8), the utterance of the test speaker is verified against each GMM of the speaker in the development set, and N speakers with best scores are selected.

- **Summation of cohort statistic** – matrices of accumulated statistics (8), (9) are initialized as the sum of all statistics from speakers in the cohort:

$$\boldsymbol{k}_{(n)i} = \sum_{s=1}^{N} \boldsymbol{k}^{s}_{(n)i}, \quad \boldsymbol{G}_{(n)i} = \sum_{s=1}^{N} \boldsymbol{G}^{s}_{(n)i}, \tag{12}$$

for each cluster n and each row i of $\boldsymbol{W}_{(n)}$. At the end of initialization, statistics of the actual test speaker are added to these initialization statistics.

B. Sufficient Statistics using Phonetic Tree

An improvement of the selection of accumulated statistics from "nearest" speakers described in the previous section is to consider also the acoustic variability within a speaker. One cannot expect to find a speaker with a voice identical to the voice of another speaker, especially if the development database is of limited size. It is more likely that the style of the pronunciation of some small part of his utterance, e.g. of a few phonemes, is same as another speaker's pronunciation of same phonemes. Hence, the motivation is to collect additional accumulated statistics not only from "nearest" speakers according to the whole test utterance, but to split up the test utterance according to its phonetic content, and find the "nearest" sound from the development set for every part (e.g. phoneme) of the test utterance. For this purpose all utterances from development speakers have to be split up too, and the statistics have to be collected beforehand.

Due to the small amount of given test data only limited amount of phonetic events will be present in the test utterance (only a few phonemes). Nevertheless, we would like to initialize also transforms related to phonemes that are not observed in the test utterance. Thus, it is useful to incorporate regression trees. Since we are dealing with phonetic events it is more appropriate to replace regression trees based on proximities in the acoustic space (6) by phonetic trees used e.g. in (9), however we will consider only three phonetic classes – vowels, consonants and non-speech events – depicted in Figure 2.

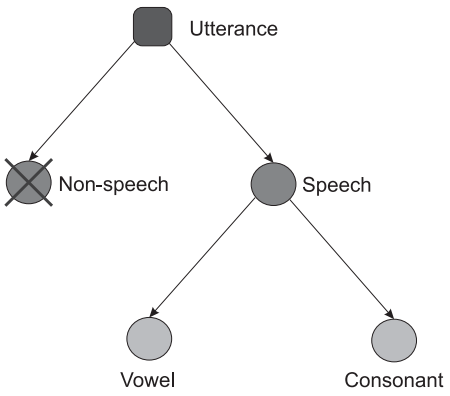

Fig. 2. An example of a phonetic tree.

The modified algorithm given in Section III-A can be now:

- **Accumulation of statistics** – for each speaker s from the development set matrices $\boldsymbol{k}^{s}_{(n)i}$ and $\boldsymbol{G}^{s}_{(n)i}$ are accumulated, but clusters $K_n, n = 1,\ldots,N$ are based on the Phonetic Regression Tree (PRT) from Figure 2. Thus, the phonetic transcription of the development utterances is performed, and feature vectors are divided to clusters in relation to the phonetic class in PRT. One GMM is trained for each speaker s and each class in PRT.
- **Selecting a cohort of phonetic events** – data from the test speaker are divided into classes in PRT according to the phonetic transcription of the test utterance. For test data in each phonetic class the N "nearest" initialization statistics are found in relation to the likelihood obtained from development GMMs. For classes with insufficient amount of test data the parent class in the regression tree is used instead, and "nearest" statistics are collected from the parent class. However, such situation is quite rare since only three phonetic classes are used – non-speech events, vowels and consonants.
- **Summation of cohort statistic** – matrices of accumulated statistics (8), (9) are initialized as the sum of all statistics from corresponding cohort of phonetic events. And at the end, statistics of the test speaker are added to these initialization statistics.

Thus, voice of each speaker is now represented not only by an average voice of "nearest" speakers, but it is in advantage piecewise composed of average phonetic events spread across many voices.

IV. Experiments

A. SpeechDat-East (SD-E) Corpus

We used the Czech part of SpeechDat-East corpus (10). Extract features are based on Mel-Frequency Cepstral Coefficients (MFCCs), 11 dimensional feature vectors were extracted each 10 ms utilizing a 32 ms hamming window, Cepstral Mean Normalization (CMN) was applied, and Δ, Δ^2 coefficients were added.

A 3 state HMM based on triphones with 2105 states in total and 8 GMM mixture components with diagonal covariances in each of the states was trained on 700 speakers with 50 sentences for each speaker (cca 5 sec. on a sentence). UBM containing 256 mixture components was trained on the same dataset, and subsequently all GMMs of individual development speakers were MAP adapted.

To test the systems performance different 200 speakers from SD-E were used with 50 sentences for each speaker, however maximum of 12 sentences was used for the adaptation. A language model based on trigrams was used in the recognition (11). The vocabulary consisted of 7000 words.

B. Adaptation Setup

In our experiments the fMLLR adaptation was utilized. Before the adaptation statistics (9) and (8) for all development speakers were precomputed (see Section III-A).

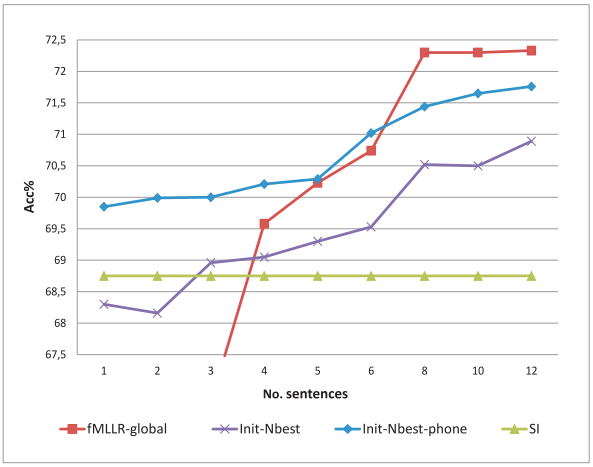

Fig. 3. Accuracy (Acc)[%] of the recognition utilizing described methods in dependence on the number of adaptation sentences.

In the case of uninitialized fMLLR only one global transformation matrix was used (all the mixture components shared the same cluster). When additional accumulated statistics from closest speaker were used for initialization multiple clusters were utilized – the clustering of model components (GMM means) was performed via a regression tree (6). In all the experiments only one iteration of fMLLR was carried out.

C. Results

All methods were tested on varying number of adaptation sentences. The graph in Figure 3 depicts results of ASR with fMLLR adaptation. Initializations are denoted as Init-Nbest (init. based on additional accumulated statistics from N-best speakers) and Init-Nbest-phon (init. based on additional accumulated statistics from N-best phonetic events). In the graph the adition information about the accuracy (68.75%) of the recognition system with the unadapted SI acoustics model is depicted. All results of adaptation are shown in the Table I.

No. of Sent.	fMLLR-global	Init-Nbest	Init-Nbest-phon
1	14.04	68.30	69.85
2	56.36	68.16	69.99
3	66.74	68.96	70.00
4	69.58	69.05	70.21
5	70.23	69.30	70.29
6	70.74	69.53	71.02
8	72.30	70.52	71.44
10	72.30	70.50	71.65
12	72.33	70.89	71.76

TABLE I
ACCURACY (ACC)[%] OF THE SYSTEM PERFORMANCE FOR FMLLR ADAPTATION AND EACH TYPE OF INITIALIZATION.

As expected, classic fMLLR (fMLLR without any initialization) performs poor on small amount of adaptation data. Init-Nbest-ph initialization outperforms the Init-Nbest initialization most for only a few sentences. After 6 sentences the contribution of initialization becomes inferior, and in real applications it would be suitable to switch the initialization off.

V. CONCLUSIONS

Presented experiments proved the contribution of the proposed initialization of fMLLR adaptation with extremely small input data sets. The initialization based on phonetic events from closest speakers turned out to be of importance. An relative improvement up to 5% was achieved in relation to the *N*-best method from Section III-A. Because the use of initialization data lowers the influence of given adaptation data, it could be suitable to weight initialization data in dependce on the amount of adatpation data. When enough data for adaptation are available, the initialization statistics are not needed.

VI. ACKNOWLEDGEMENTS

This research was supported by the Ministry of Culture Czech Republic, project No. DF12P01OVV022.

REFERENCES

C. J. Leggeter, and P. C. Woodland, “Maximum Likelihood Linear Regression for Speaker Adaption of Continuous Density Hidden Markov Models”, Computer Speech and Language, vol. 9, pp. 171-185, 1995.

M.J.F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition”, Computer Speech and Language, vol. 12, pp. 75–98, 1997.

D. Povey, and K. Yao, “A Basis Representation of Constrained MLLR Transforms for Robust Adaptation”, Computer Speech & Language, vol. 26, pp. 35–51, 2012.

Y. Li, H. Erdogan, T. Gao, and E. Marcheret, “Incremental on-line feature space MLLR adaptation for telephony speech recognition”, 7th International Conference on Spoken Language Processing, pp. 1417–1420, 2002.

Z. Zajíc, L. Machlica, and L. Müller, “Initialization of fMLLR with Sufficient Statistics from Similar Speakers”, Lecture Notes in Computer Science, vol. 6836, pp. 187–194, 2011.

M. J. F. Gales, “The Generation and use of Regression class Trees for MLLR Adaptation”, Techreport Cambridge University Engineering Department, 1996.

D. Povey, and G. Saon, “Feature and model space speaker adaptation with full covariance Gaussians, Interspeech 06 - ICSLP, pp. 1145–1148, 2006.

D. A. Reynolds, T. F. Quatieri, and R. D. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models”, Digital Signal Processing, vol. 10, pp. 19–41, 2000.

A. Stolcke, S. S. Kajarekar, L. Ferrer, and E. Shriberg, “Speaker recognition with session variability normalization based on MLLR adaptation transforms”, IEEE International Conference on Spoken Language Processing, vol. 15, pp. 1987–1998, 2007.

P. Pollak, et al., “SpeechDat(E) - Eastern European Telephone Speech Databases”, XLDB - Very Large Telephone Speech Databases (ELRA), 2000.

A. Pražák, J. Psutka, J. Hoidekr, et al., “Automatic online subtitling of the Czech parliament meetings”, Lecture Notes in Artificial Intelligence, vol. 4188, pp. 501–508, 2006.