Telecommun Syst
DOI 10.1007/s11235-011-9660-8

Building accurate and robust HMM models for practical ASR systems

Juraj Kačur · Gregor Rozinaj

Abstract In this article the relevant training aspects for building robust and accurate HMM models for large vocabulary recognition system are discussed and adjusted, namely: speech features, training steps, and the tying options for context dependent (CD) phonemes. As the basis for building HMM models the well known MASPER training scheme is assumed. First the incorporation of the voicing information and its effect on the classical extraction methods like MFCC and PLP will be shown together with the derivative features, where the relative error reductions are up to 50%. Next the suggested enhancement of the standard training procedure by introducing garbled speech models will be presented and tested on real data. As it will be shown it brings more than a 5% drop in the error rate. Finally, the options for tying states of CD phonemes using decision trees and phoneme classification will be adjusted, tested, and explained.

Keywords Speech recognition · Hidden Markov models · Speech features · Model training · MASPER

1 Introduction

For a couple of decades there has been a great effort spent in building and employing automatic speech recognition (ASR) systems in areas like information retrieval systems, dialog systems, etc., but only as the technology has evolved to some stage other applications like dictation systems or even automatic transcription of natural speech [1] are emerging. These advanced systems should be able to operate on a real time base, must be speaker independent, achieving high accuracy even in the presence of additive and convolution noises, changing environments, and support dictionaries containing several hundreds of thousands of words.

These strict requirements can be currently met by HMM models of tied CD phonemes with multiple Gaussian mixtures, which is a technique known from the 60ties [2]. As this statistical concept is mathematically tractable it, unfortunately, doesn't completely reflect the physical underlying process. Therefore, soon after its creation there have been many attempts to alleviate that. Nowadays the classical concept of HMM has evolved into areas like hybrid solutions with neural networks, utilization of different than maximum likelihood (ML) or maximum a posteriori probability (MAP) training strategies that minimize recognition errors by the means of corrective training, maximizing mutual information [3] or by constructing large margin HMMs [4]. Furthermore, a few methods have been designed and tested aiming to suppress the first order Markovian restriction by e.g. explicitly modelling the time duration (Levinson, 1986), splitting states into more complex structures [5], using double [6] or multilayer structures of HMM. Last but by no means least issue is the construction of robust and accurate feature extraction method. Again this matter is not fully solved and various popular features and techniques exist like: Mel frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP) features, concept of RASTA filtering [7], time frequency filtering (TIFFING) [8], Gammatone cepstral coefficients (GTCC) [9], zero-crossing peak amplitude (ZCPA) [10], etc.

Even despite the huge variety of advanced solutions in ASR systems many of them are either not general enough (specially designed to certain environment) or are rather impractical for the real-life employment. Thus, in the present time most of the practically employed systems are based on

J. Kačur (✉) · G. Rozinaj
Faculty of Electrical Engineering and Information Technology,
Slovak University of Technology, Bratislava, Slovakia
e-mail: kacur@ktl.elf.stuba.sk

Published online: 06 October 2011

continuous context independent (CI) or tied CD HMM models of phonemes with multiple Gaussian mixtures trained by ML or MAP criteria. As there is no analytical solution for the optimal setting of HMM parameters given the real data, there must be a training process employed which is an iterative one [3]. Unfortunately, using it, there is no guarantee of reaching the global maxima [11], thus lot of effort is paid to the training phase in which many procedures are selectively applied in stages. Thus only mature and complex systems allow convenient, efficient and flexible training of HMM models, where the most famous are HTK and SPHINX. These systems are looked at as standard tools for building robust and accurate models for large vocabulary, practical systems.

This article discuses and explains major stages of building speaker independent continuous density HMM (CDHMM) models using the professional database MOBILDAT-SK [12] and the standard training scheme called MASPER [13]. The rest of the article is organized as follows. First the standard training method will be used with several base speech feature extraction methods, namely MFCC and PLP and a couple of auxiliary parameters, mostly dynamic ones. Further the measure of voicing which plays a role in differentiating some pairs of consonants will be added as well as the pitch itself. Each setting and feature will be tested and its merit numerically assessed and compared to the original one. In the second section the focus will be on the training process itself, where as the basis the reference recognition system REFREC [14] or its multilingual version MASPER will be used and introduced in brief. However the core part would be on presenting the enhancement to the standard training procedure by incorporating the background models of garbled speech. Several structures of those models will be designed and tested as well as the ways how they are to be optimally trained. The third part deals with the process of tying HMM states of CD phonemes using both the language information (classification of phonemes) and the statistical similarities of the data. Setting optimal tying options, understanding the underlying physical process and its effect on the right balance between the accuracy and generality may bring additional increase of the overall accuracy. Next, in brief, the training and testing environments (executed tests) will be presented along with the professional database. The article is concluded by summarizing results and findings. Therefore the presented article should give you an insight into how to adjust and build both robust and accurate HMM models using standard methods and systems on the professional database. Further, it should suggest what may be and what probably is not so relevant in building HMM models for practical applications, i.e. which part the designer should be particularly careful with.

2 Feature extraction methods and their performance

One of the first steps in the design of an ASR system is to decide which feature extraction technique to use. At the beginning it should be noted that this task is not yet completely solved and a lot of effort is still going on in this area. The aim is to simulate the auditory system of humans, mathematically describe it, simplify for practical handling and optionally adapt it for a correct and simple use with the selected types of classification methods.

A good feature should be sensitive to differences in sounds that are perceived as different in humans and should be “deaf” to those which are unheeded by our auditory system. It was found [15] that the following differences are audible: different location of formants in the spectra, different widths of formants and that the intensity of signals is perceived non-linearly. On the other hand, following aspects do not play a role in perceiving differences: overall tilt of the spectra like, filtering out frequencies lying under the first formant frequency, removing frequencies above the 3rd format frequency, and a narrow band stop filtering.

Furthermore, features should be insensitive to additive and convolutional noises or at least they should represent them in such a way that these distortions are easy to locate and suppress in the feature space. Finally, when using CDHMM models it is required for the feasibility purposes that the elements of feature vectors should be linearly independent so that a single diagonal covariance matrix can be used. Unfortunately, there is no feature that would ideally incorporate all the requirements mentioned above.

Many basic speech features have been designed so far, but currently MFCC and PLP [16] are the most widely used in CDHMM ASR systems. In most cases these basic features aim to mimic the static part of the spectra as it is perceived by humans. Apart from the static features it was soon discovered that the changes in the time [17] represented by delta and acceleration parameters play an important role in modelling the evolution of speech. This notion was further evolved by introducing the concept of modulation frequency and RASTA filtering [7]. This is particularly important when using HMMs as they lack the natural time duration modelling capability (geometric time distribution in a single state). Overall energy or zero cepstral coefficients with their derivations also carry valuable discriminative information thus most of the systems use them.

More details on both MFCC and PLP methods can be found elsewhere [9, 16, 18], but in the following let us mention some basic facts and achieved results both for basic features alone and with auxiliary coefficients.

2.1 MFCC vs. PLP

MFCC and PLP both represent some kind of cepstra and thus are better in dealing with convolutional noises. However, it was reported that some times in lower SNRs they

Table 1 WERs and relative improvements for MFCC, PLP and auxiliary features averaged over different HMMs for digit strings and application words tests

	Static	Relative Improvements related to previous step			
	WER [%]	C0 [%]	Delta [%]	Acceleration [%]	Cepstral mean subtraction [%]
PLP	40.3	11.1	61.47	52.1	23.27
MFCC	43.6	20.64	61.36	48.15	52.66

are outperformed by other methods, e.g. TIFFING [8], J-RASTA [19], ZCPA [10], etc. as they nonlinearly couple signals with additive noises. This well known fact of a rapid deterioration in accuracy for lower SNRs that is not solely related only to PLP or MFCC is compensated by matching training and employment environments, thus huge professional databases are recorded in real environments. On the other hand, in the clear environment (above 30 dB) MFCC and PLP provide one of the best results [10].

The computation of MFCC undergoes following processing.The speech signal is first modified by HP so-called preemphasis filter to suppress the low pass filtering character of the speech given by the lip radiation to the open space. Prior to the FFT computation a Hamming window is applied and the frequency in Hz is warped into the Mel scale to mimic the critical bands over different frequencies. Next, equally spaced triangular windows with 50% overlap are applied to simulate a filter bank. Finally the logarithm function and DCT transform are applied that produce a static feature frame. The logarithm not only acts as a tool to produce cepstrum (real one) but suppress the high-value intensity in favor of low intensities as the human auditory system does. In addition, zero cepstral coefficient is used as well to estimate the overall log energy.

Unlike MFCC the original process of PLP calculation follows these steps: calculation of FFT that is proceeded by Hamming windowing, frequency warping into Bark scale, smoothing the bark-scaled frequency spectra by a window simulating critical bands effect of the auditory system, sampling the smoothed bark spectrum in approx. 1 bark intervals to simulate the filter bank, equal loudness weighting of the sampled frequencies which approximates the hearing sensitivity, transformation of energies into loudness by powering each frequency magnitude to 0.33, calculating the linear prediction (LP) coefficients from the warped and modified spectra (all pole model of the speech production), finally cepstral LP coefficients are derived from LPC as if the logarithm and the inverse FFT were calculated.

In both MFCC and PLP cases, the DCT or FFT transform applied in the last step of the computation process minimize the correlation between elements and thus justifies the usage of diagonal covariance matrices. Furthermore, to take the full advantage of cepstral coefficients, usually a cepstral mean subtraction is applied in order to suppress possible distortions inflicted by various transmission channels or recording devices. At the end we shall not forget about the liftering of a cepstra in order to emphasise its middle part so that the most relevant shapes of spectra for recognition purposes would be amplified (lower-index coefficients approximate the overall spectral tilt and the higher-index coefficients reflect the details and are prone to contain noise) [15]. Well, this appealing option has no real meaning when using CDHMM and Gaussian mixtures with diagonal covariance matrices only. In this case it is simply to show that the liftering operation would be completely canceled out when computing Gaussian pdf.

All the above-mentioned features and auxiliary settings were tested and evaluated on he MOBILDAT-SK database in terms of the recognition error. Two tests were done on the test set portion of the database: digit strings, and application words whose results were averaged. The training was based on the MASPER training procedure (will be presented later in the text) using the HTK system. In Table 1 there are shown averaged word error rates (WER) for PLP and MFCC features as scored in the application words and digit string tests. The relative improvements achieved by: adding zero cepstral coefficient, including delta and acceleration coefficients, and applying cepstral mean subtraction, are also shown. These results were calculated and averaged over different HMM models i.e. CI and tied CD phoneme models with multiple Gaussian mixtures.

From these tests one can induce that slightly better results are obtained by PLP method in both cases, once regarding only the static features (43.6% vs. 40.3% of WER in favor for PLP) and the other time using all abovementioned static and dynamic auxiliary parameters and modifications (5.24% vs. 5.08% of WER in favor for PLP). Further, we investigated step by step the significance of auxiliary features and modification techniques. First let us begin with the zero cepstral coefficient (C0), where its incorporation brought relative improvements over basic PLP (11.1%) and MFCC (20.64%) vectors. As it can be seen the improvement is much more relevant for MFCC, thus we can interpret this result as the PLP provides more complex representation of the static speech frame than MFCC (from the point of view of recognition accuracy), because the additional information was not so beneficial. Next the inclusion of delta coefficients disclosed that their incorporation brought down the averaged error relatively by 61.36% for MFCC and 61.47% for PLP (related to basic vectors plus C0). If this absolute drop is further transformed to the relative drop calculated over a single difference coefficient (if all are equally important), it shows that one delta coefficient on average causes a 4.72% WER drop for MFCC and 4.73% for PLP. Next, the acceleration coefficients were tested, and their inclusion resulted in a 48.15% drop of WER for MFCC and 52.1% drop for PLP relative to the previous setting (basic vector + C0 + delta). Again, the incorporation of dynamic (acceleration) coefficients was more beneficial for PLP. If these absolute drops in WER are calculated for a single acceleration coefficient, it was found that one such coefficient causes on average a 3.7% drop of WER for MFCC and a 4% for PLP. Finally, the cepstral mean subtraction was tested for both methods where it brought substantially improved results on average by 52.66% for MFCC and 23.27% for PLP. As it can be seen the benefit of this operation is tremendous for MFCC comparing to PLP. That reveals the PLP is less sensitive to the cepstral mean subtraction, probably, because it uses non linear operations (0.33 root of the power, calculation of the all pole spectra) applied prior to the signal is transformed by the logarithm and before the cepstral features are calculated. Interestingly enough, both dynamic features caused to be more significant for PLP than for MFCC in relative numbers, however, for the additional C0 (static feature) this was just the opposite. All this may suggest that PLP itself is better in extracting static features for speech recognition as the information contained in C0 and cepstral mean subtraction are not so vital, unlike MFCC.

2.2 Voicing and the pitch as the speech features

As it was shown in the previous paragraph apart of the base static features that aim to estimate the magnitude-modified and frequency-warped spectra, the dynamic features reflecting time evolution, like delta and acceleration coefficients proved to be rather valuable as well. As a consequence of their construction and aim, those parameters eliminate any voicing evidence contained in the signal which still carries some discriminative information that may play a role in the recognition accuracy. Namely, this information is vital to discriminate between some pairs of voiced and unvoiced consonants (t–d, etc.). To alleviate this drawback we substituted the least significant static features with such kind of information derived from the average magnitude difference function (AMDF). To evolve this concept further the pitch was tested in the separate experiments as well. In order to verify and assess the merit of such a modification, series of experiments were executed using the professional database and a training scheme for building robust HMM models.

As it was already outlined the concept of incorporating some parameters assessing the voicing of a particular signal may bring additional discriminative information into the recognition process. For example in Slovak, one classification of consonant is according whether they exist in voiced/unvoiced pairs. In the group where pairs exist each consonant is matched into pair according to the mechanism how they are produced and perceived. In the group of pairs there is always an unvoiced consonant that is matched to the voiced one. The only difference in their production is the absence of vocal chord activity that obviously can not be observed after PLP or MFCC processing. Some typical pairs of voiced and unvoiced consonants are: z in the word zero and s as appears in the word sympathy, p (peace) and b (bees), d (divide) and t (tick), etc. As it can be seen, distinguishing between them may be crucial, thus such information can be potentially beneficial. On the other hand, it must be said that there are many cases (at least in Slovak) where in the real conversation the voiced and paired consonant may take the form of unvoiced one and vice versa. Thus these two contradictory effects must be tested and assessed, to see which one is prevailing.

As it comes to the selection of proper method estimating the voicing degree in a signal, there are more methods to do it as well as to detect the periodicity as a side product. These algorithms are ranging from the simple algorithms in the time domain like: AMDF, and autocorrelation, through spectral ones like harmonic spectra and real cepstrum to methods operating on the residual signal after the inverse filtering. A good method should be accurate, robust against additive noise, easy to compute and the outcome should be easy to interpret and should be gain invariant. In our experiments we opted for AMDF method as it provides good results obtained even in lower SNRs, has a simple and fast implementation, its output has straightforward representation, the by product is the detected pitch and moreover it is the base for more complex methods like YIN [20]. The AMDF function is defined as follows:

$$f_i(k) = \sum_{n=0}^{n<N} |s(i \cdot N + n) - s(i \cdot N + n + k)|,$$
$$k \in \langle T_{\min}, T_{\max} \rangle, \qquad (1)$$

where s is a signal, N is length of the tested block i and $T_{\min}$ and $T_{\max}$ are the minimal and maximal pitch periods. Then as a measure of voicing the minimal value of AMDF found within the eligible ranges for pitch can be used. Further, to suppress its dependence upon magnifying constant, usually a ratio to its maximal or averaged value is computed as follows:

$$voicing_i = \frac{\frac{1}{T_{\max}-T_{\min}} \sum_{k=T_{\min}}^{k<T_{\max}} f_i(k)}{\min_{T_{\min} \le k < T_{\max}} (f_i(k)).} \qquad (2)$$

Using this definition the voicing measure is in the range $\langle 1, \infty)$. Therefore, its interpretation is as follows: for mathematically periodic signals it produces infinite value—voicing, whereas random signals with no period would ex-

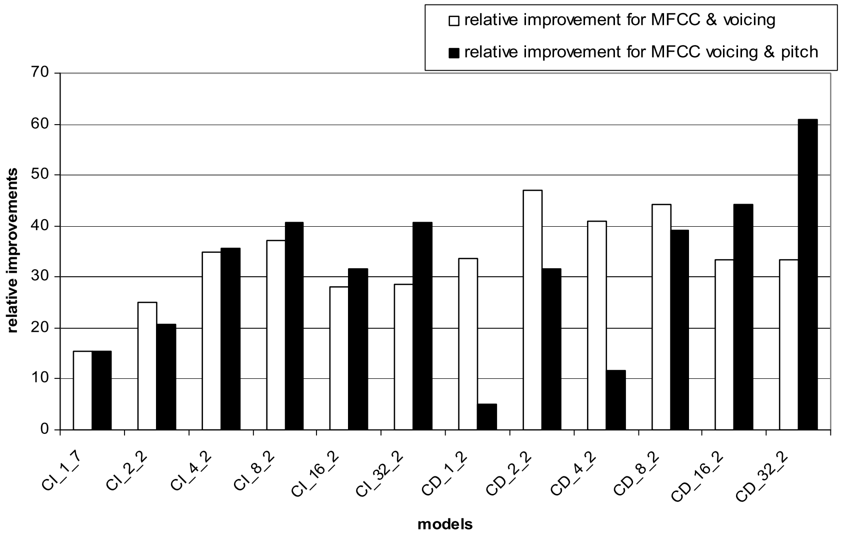

Fig. 1 Relative improvements for MFCC, various HMM models (CI and CD with 1 to 32 mixtures each) and application words test achieved by incorporating voicing and pitch measures

hibit values close to 1. Furthermore, the location of the minima can represent the estimated pitch; however certain precaution and logic should be introduced in order not to detect longer periods—usually integer multipliers of the real one.

To assess its effectiveness two basic tests were executed upon all models and the examined feature extraction methods, i.e. MFCC and PLP both in combination with the voicing and pitch estimates. All experiments were accomplished on the test part of the MOBILDAT-SK database that amounts to 220 speakers. The two basic tests were: digits strings that can contain arbitrary number of digits in any string and the application words test that equals to the recognition of isolated words. Although the digits string test exhibits higher perplexity and therefore provides higher errors it uses only very limited set of CD phonemes. On the other hand, the application words test contains greater variety of words, CI and CD phonemes and therefore provides more objective insight about how good all the models were trained.

To assess the merits of incorporating the voicing measure and the pitch into the recognition process, achieved results are related to the original ones (MFCC and PLP) via the relative improvement unit defined as follows:

$$relative\_improvement = \frac{WER_{orig} - WER_{mod}}{WER_{orig}} 100. \qquad (3)$$

Further it should be noted that the new features were not simply added to the original vector but instead they replaced its least significant elements (we believed according to the theory they were the last PLP and MFCC coefficients). This allowed us to maintain the same vector size during all the experiments and thus made the comparison more objective.

As the PLP and MFCC features performed in a slightly different way the results will be given separately for both features. Finally, to save up the space only results for the application words test are shown in the following pictures, as similar behavior was observed in the case of digits strings, however due to the higher perplexity higher errors were achieved.

In Fig. 1 there are depicted the relative improvements for MFCC achieved by introducing the voicing measure and the pitch feature in the case of application words test. The same results for PLP are shown in Fig. 2. In all the cases it is clear that better results were observed by the inclusion of suggested features, however the replacement of the least significant basic features is more relevant for MFCC where on average an 24.51% improvement was recorded whereas for PLP it was only 19.96% (average for application words and digit strings tests). Further, it is noticeable that the pitch was more successful (black bars) in connection with MFCC features and that its incorporation tended to be more beneficial in the case of more complex models (with higher number of mixtures) for both PLP and MFCC parameters.

To conclude the discussion on the speech feature extractions and the voicing measure, let us recap and explain the observed results in the following points.

Incorporating the voicing coefficient defined in (2) and (3) that assessing the level of similarity between possible periods of voiced signals turned out to be very effective; on average a 24.51% improvement was observed for MFCC and 19.96% for PLP. This feature may provide necessary discriminative information between paired consonants (voiced and unvoiced, like: s and z, p and b, t and d, etc.).

The inclusion of pitch itself by replacing the least significant PLP or MFCC elements was despite its dispersion and non-lexical-nature beneficial in both cases, where for MFCC the averaged improvement reached 21.68% and for PLP it was only 17.6%. Those relative “improvements” are in fact degradation in the accuracy compared to the voicing measure and not to the original basic features; then the WER increased by almost 3% for MFCC and more than 2%

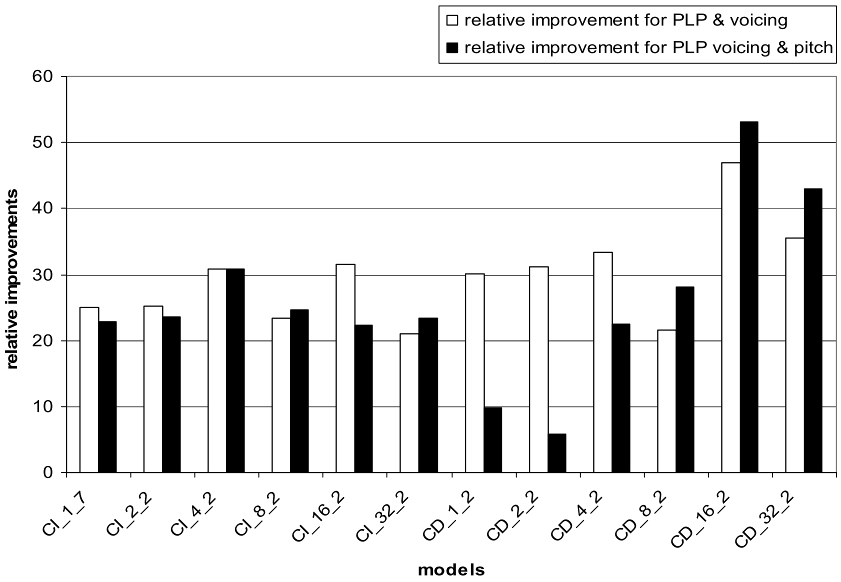

Fig. 2 Relative improvements for PLP, various HMM models (CI and CD with 1 to 32 mixtures each) and application words test achieved by incorporating voicing and pitch measures

for PLP. On the other hand the pitch information is more beneficial for more complex models even in the PLP case; however benefits were relatively minor. This phenomenon can be explained by better description capabilities of more complex HMM models that are more equipped to cope with the wider ranges of possible pitch values.

The results with voicing and pitch parameters are in line with those mentioned in Table 1, where it was shown that PLP static features were more effective in representing the speech for recognition purposes. Thus the replacement of the least significant features by voicing and pitch parameters was more relevant for MFCC extraction method in all the cases; more details can be found in [21].

3 Training process for HMM models

There have been several projects aimed at setting a general and reference framework for constructing robust and accurate ASR system for real life applications [22] using databases recorded over fixed or mobile telephone networks. Currently the common methods or concepts are the REFREC (versions 0.95 and 0.96) [14, 23] and its multilingual counterpart MASPER [13]. They are tailored to operate on the SPEECHDAT [24] or MOBILDAT [12] databases, as these act as standards for recorded data files, annotation conventions, labels files, etc., and there are many of them (more than couple of dozens) in different languages. Furthermore, both training methods are based on the advanced, wide spread and well documented HTK toolkit [25] and were designed as a part of COST 249 initiative.

The main concept of REFREC 0.95 recognition system is based on that one presented in HTK tool [25], however enhanced to serve for multilingual purposes. During the course of the run it produces following models: flat start (no time alignment) monophones with 1 up to 32 Gaussian mixtures, time aligned monophones (1 to 32 mixtures), triphones (with only one Gaussian) and tied models of triphones with 1 to 32 Gaussian mixtures. As a part of the training there are 3 unified small vocabulary tests provided for all models involving application words, single digits and looped digits.

In the following a brief introduction to the training process of REFREC and MASPER is given. The speech recordings are parameterized to the MFCC vectors with 13 static coefficients including C_0. Auxiliary dynamical features like delta and double delta coefficients are appended during the course of training to make up the total vector length of 39. The data preparation process goes on by: gathering and transforming the descriptions files into master lab file [25], modifying dictionary entries, selecting the training, testing and development testing sessions, producing phone lists, phoneme mapping, etc. Further, individual utterances that are in a way damaged by the presence of intermittent noise (int), unintelligible speech ($**$), mispronunciation ($*$), truncation ($\sim$) and filled pauses (fil) are removed, while other markers are ignored. Then the training starts by flat start adjustment of all Gaussians to the averaged mean and variance of observed vectors. This initialization phase and first cycles of embedded trainings are executed over the phonetically rich utterances with creation of silence (SIL) and short pause (SP) models. Then a Viterbi realignment is done throughout the whole database and further training cycles with gradually increasing number of Gaussian mixtures up to 32 are performed. These final models are used to do the time alignment of all utterances so that a single model training of monophone models can be done in the second stage of the training (modification to 0.94 version). These aligned models are built and enhanced in the similar fashion as previously, but from the time aligned data. Monophones derived in this stage with 1 Gaussian mixture are used for cloning context dependent phonemes (triphones). As only word internal triphones are regarded, the remaining parts are modeled by biphones. Triphones are first trained in 2 cycles of embedded training and then tied to increase their robustness and to cut down on their size. This is done via the clustering decision tree algorithm which is based on the training data, but still follows particular rules of the language set by experts. This is done by generating questions about the left and right context utilizing provided classes of phonemes aiming at increasing the probability of the modeled data. Furthermore, this approach enables to create structures for synthesizing unseen triphones in the training set.

In the 0.96 versions some non-speech acoustic events were modeled rather than the whole affected utterances would have been excluded, which usually ended up in relatively large drop in the training data. Thus utterances degraded by intermittent noise (Int) and filled pauses (Fil) are preserved in the training set and models of speaker produced noise (Spk) and hesitation model (Fil) are introduced and trained. However Int markers are left unheeded in the training and all these amendments brought overall improvements as documented in [14].

To enable an effective design of multilingual and cross-lingual ASR systems (main goal) some further modification must have been done to the REFREC 0.96, which resulted in MASPER procedure [13]. These changes are as follows: cepstral mean normalization (channel equalization), modifications to the parameters of tree based clustering, production of training statistics, and distributed computation.

As can be seen the modified REFREC 0.96 or MASPER is an advanced procedure for building mono, multi or cross-lingual HMM models. However, we discovered some deficiency in handling with the training data, i.e. the removal of all utterances contaminated with truncated, mispronounced and unintelligible speech. Thus in the following the modification to the MASPER procedure aiming to model the damaged parts while preserving useful information for the training purposes will be discussed.

3.1 Tested models for Garbled speech and the statistic of training data

One of the possible ways to improve the performance of any ASR system based on HMM models is to make these models in the training phase more robust while preserving their accuracy. These two requirements contradict to each other and fulfilling both ends up in the increasing number of the training data which is difficult and expensive to gather. REFREC 0.96 fixed this comparing to its 0.95 version by allowing some damaged utterances to be used in the training process and by doing so it improved the overall results [14]. However, there are still some recordings excluded because of serious damages. We proposed methods to solve this problem and save these recordings for the training process. First let us recap some facts.

For the training purposes there are 44000 speech files available in MOBILDAT-SK database. However, training speech files that contain mispronounced, truncated and unintelligible speech or speech contaminated with frame loss or fading in the mobile network (marked by %), are rejected from the MASPER training process. It is so because for these files it is not possible to create correct phoneme-level transcriptions, as they can be clipped from the sentence, and further, the transcription of any sentence does not contain any kind of time markers. MOBILDAT-SK was found to contain 3096 of such speech files. These are doomed to be rejected although they can still contain noticeable count of phonemes that could be used in the training process, and more than 26% of all rejected speech files are items containing phonetically rich sentences.

Using the phone level transcriptions created by MASPER procedure, the phoneme analysis of MOBILDAT-SK has been performed. After the rejections of corrupted speech files there were in total 955611 instances of all phonemes, of course, except sil, sp, fil and spk models. The same analysis applied just to the rejected speech files has brought further 89018 realizations of usable phonemes, which amounts to the 9.32% average increase of usable phoneme instances. More detailed statistic regarding the recordings, phonemes and triphones found by MASPER and modified MASPER procedures on MOBILDAT-SK is summarized in Table 2.

Furthermore, the number of phoneme instances for each phoneme is displayed in Fig. 3. The modified MASPER procedure preserves useful data from damaged speech for the training purposes. It does so by using unified model of the garbled speech so the damaged recordings can be modeled and thus do not need to be rejected. The corrupted words in these utterances are not expanded to the sequence of phonemes, but instead they are mapped to a new unified model of garbled speech while the rest of sentence can be processed in the same way as in the classical procedure. The new model is added to the phoneme list (context independent and serves as word break) and trained together with other models. However its structure must be more complex as it should map words of variable lengths spoken by various speakers in different environments.

The reasons mentioned above for the great complexity of the new model imply that the structure contains all possible forward and backward transitions except transition between non-emitting states themselves. That means the unified model is ergodic, as it theoretically allows finding the best assignment of each observation to any state regardless

Table 2 Statistics of utilized phonemes, triphones and recordings in MOBILDAT SK by MASPER and modified MASPER

Statistics of the database	MASPER	Modified MASPER	Absolute increase	Relative increase
recordings	40861	43957	3096	7.58%
phonemes	51	51	0	0%
triphones	10567	10630	63	0.60%
instances of phonemes	955611	1044629	89018	9.32%
average number of instances per a triphone	∼90.4	∼98.27	∼7.84	∼8.7%

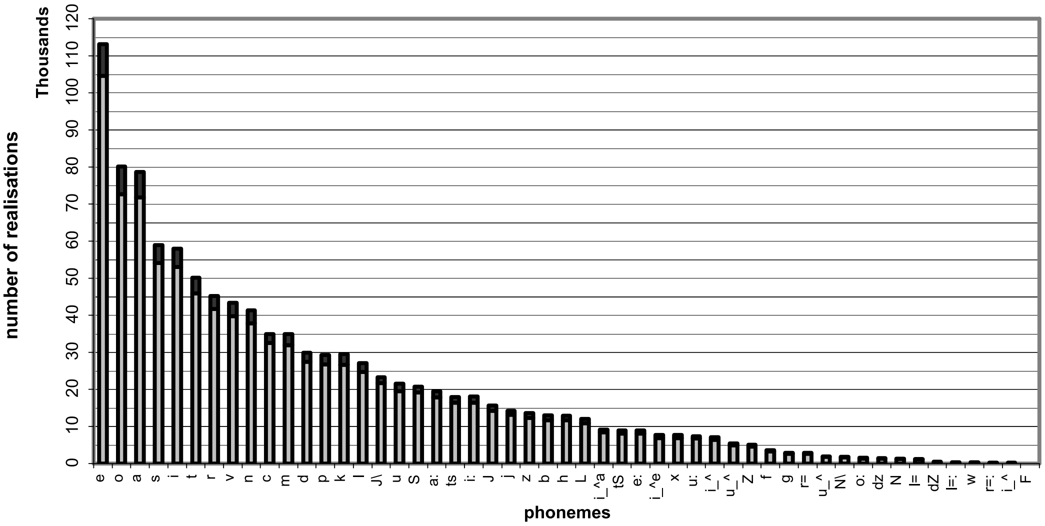

Fig. 3 Number of instances for each phoneme in MOBILDAT-SK processed by MASPER and modified MASPER

of state's sequential number. On the other hand, this structure increases the number of possible paths (estimated parameters) through the model, and thus can be difficult to train. Thus several experiments have been performed in order to compare the results achieved by using different structures of the garbled speech model and its enhancements during the training process to find the best combination of the model structure and its training for the given speech database. Individual structures differed in the number of emitting states while the final ergodic transition matrix was used.

The simplest structure contains only one emitting state which describes general characteristics of the garbled speech signal itself. Nevertheless the number of Gaussian mixtures grows up to 32 during the MASPER training, which then leads to the finer probability space subdivision. This structure has the least adjustable parameters, and so can be robustly estimated. On the contrary the most complex structure of the garbled speech we tested consisted of 5 emitting states. In this case all possible transitions between emitting states and transitions between non-emitting and emitting states were allowed and reestimated during the training process. This model facilitates the finest probability space subdivision, but with 32 Gaussian mixes the number of adjustable parameters grows up significantly. Therefore the third structure consisting of 3 emitting states was tested and is depicted in Fig. 4. This number of states is typically used for all other models in the training including models of noises and background. The usage of ergodic transition matrix allows mapping of statistically different parts of speech to the different states. Further details on tested and trained BH model structures can be found in [26].

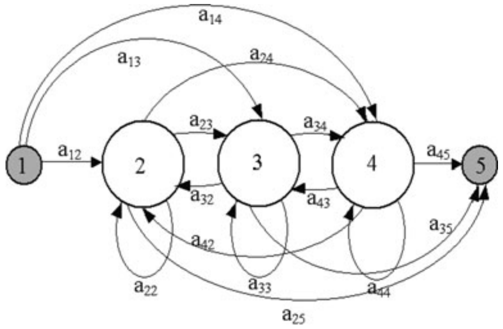

Fig. 4 The structure of 3 state model (BH) of garbled speech

3.2 Modified training procedure

The original training procedure is modified in the stage of preprocessing as well as in both training phases (CI and CD

Table 3 Tested structures and enhancement methods of BH models

Initial structure	Method of enhancement	Final structure
Left-right, start and end in the first and last emitting states	Addition of all backward connection, no T model	Ergodic
Ergodic	No	Ergodic
Left right, all forward connections	Addition of backward connections	Ergodic
Left right, all forward connections	Addition of backward connections, single model training of BH in second stage	Ergodic
Left right, all forward connections	No	Ergodic

models). In the preprocessing stage the feature extraction is made upon the training speech files and the generation of transcription files includes damaged speech as well. The name of the newly introduced garbled speech model BH (black hole—any speech can be attracted to it) is added to the dictionary and the previously generated transcriptions are converted into label files where all corrupted words are replaced with the new model.

For each tested structure of the BH model several enhancements during the training process were carried out. First kind of improvement is common to all trained models and it is the gradual upgrade of Gaussian mixtures. This is performed during the training process for all other models at the same time. The second enhancement deals with the modification to the transition probability matrix. In these cases the transition matrix entering the initialization process is less complex (strictly left-right) than the transition matrix used in further training, in other words, some transitions are prohibited at the stage of initialization as the precision is of less importance than the robustness. This enhancement is of course applicable only to BH models with more than 1 emitting state.

Enhancements to the transition matrix are performed in both stages of the training. In the first stage this happens after the "flat-start" initialization and the following two embedded reestimation cycles that are performed on the bootstrap part of the database. After that the prohibited transitions are allowed and are set to roughly similar values. From this point onwards the monophone models are trained and upgraded together with BH model as defined in MASPER. The same procedure takes place in the second stage (time marks exist and single model training is used) of the training just after the Viterbi initialization of phonemes. For ergodic BH models, the initialization with Viterbi algorithm is omitted and the BH model from the previous stage is taken. This problem does not exist in the initialization phase (cloning) of triphones, so the ergodic BH models can be trained and enhanced together with other triphones.

Three different transition matrices were tested in the initialization of the garbled speech models with higher number of states. The first one was an ergodic model just to disclose the advantages or disadvantages of the gradual enhancements of models. Thus in this case no further improvements were applied to the structure, but in order to avoid problems with uniform segmentation using Viterbi initialization [25], the model entering the second stage is copied from the last set of monophone models with one Gaussian mixture obtained in the first stage. In other tested cases the Viterbi initialization of the BH models in the second stage was used, while their structure was left-right. The second tested transition matrix contains all forward transitions except the transition between non-emitting states initialized to uniform values. All backward transitions were suppressed. The third one was a diagonal matrix allowing only transition to the next state or persisting in the current state. In Table 3 a summary of tested initial and final structures of BH models together with their methods of gradual enhancements are provided.

All the suggested BH models and their methods of gradual enhancement were tested using classical MASPER test procedures: application words, digits, and digit strings. However the most discriminative and thus more relevant results from the perplexity point of view were digit strings. In Fig. 5 there are depicted WER values achieved by the best BH models in digit strings test for both monophone and tied triphone models. The garbled speech model used in this case consists of 5 emitting states with its initial transition matrix containing all forward transitions. Other models and methods showed slightly worse results. This figure also compares the achieved results with corresponding results produced by the standard MASPER procedure. As it can be seen overall improvements were achieved which is more significant for more complex models of CD phonemes, for which the trade-of between robustness and accuracy is more critical. For these models with different complexity at least a 5% relative improvement in WER was observed. However, the negative effect of the modified MASPER procedure is that it naturally lasts longer; slightly less than an 11% time increase was observed.

Besides the time-consumption drawback (BH is trained in all stages), it was observed that there is almost no effect for CI phonemes as originally there is enough data for training 51 phonemes in uncorrupted recordings. On the contrary, the damaged recordings were also involved in the training

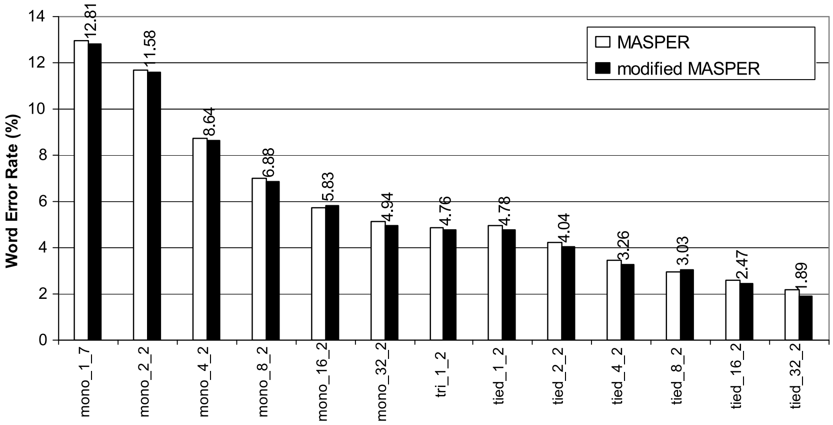

Fig. 5 An accuracy comparison of the modified MASPER procedure to the original one based on SVWL test for mono and tied HMM models

together with not well trained BH model (was not very effective as a patch covering the garbled speech) which actually resulted on average in a slightly worse results for CI phonemes. Thus a new method for the training of BH model was suggested and tested. The idea is, instead of testing new structures of BH models which provided only minor differences a new strategy was adopted. A 3 state BH model is not included in the training of CI models in both stages as it has no effect besides the increased training time. Instead, it is trained separately using the corrupted recordings only and the CI models that were already and properly trained on "healthy" recordings. The BH training undergoes gradual enhancements and extra iteration cycles can be added as the overtraining phenomenon is not a problem for BH model because it will be applied only to the training phase and seen (damaged) recordings, so the robustness can give a way to the accuracy. This amendment saves time with no effect on CI models and a more accurate BH model can be obtained. Then such complex model (ergodic with 32 Gaussian mixtures) is used as a patch when CD or tied CD phoneme models are gradually trained and enhanced including also the damaged recordings. This altered strategy besides reducing the training time brought on average a 2.65% relative improvement for all models, 3.2% for CI models and a 0.77% improvement for CD models referring to the previously suggested BH model strategy. The positive effect on CI models is caused because the low complexity models in the early stages were trained on corrupted speech together with BH models that were not complex enough to fit the garbled pronunciations. In fact, for those low complexity CI models the original introduction of the BH models caused on average the accuracy degradation.

4 Tying options for tied CD phonemes

Another training related issue that is partially language and database specific is the tying process of CD phonemes. This process merges linguistically and/or in feature space similar states. The goal is to increase the robustness of specialized CD models. The decision tree-based clustering uses language questions derived from the predefined classification of phonemes which is language specific. These questions are asked about the left and right context of a CD phoneme and based on each question the original cluster is divided according to the answer. Resulting two clusters are assumed to be better in modeling the data and thus this separation causes the increase of the log likelihood of the data. Only those clusters are left which recorded the highest increase and the process is stopped if the increase is less than the predefined minimal log likelihood. To prevent forming clusters with insufficient data the minimal state occupancy is set (occupancy count). The greatest advantage of this method is the existence of decision trees that can be later used to synthesized unseen CD phonemes [14]. Altogether 40 different classes of Slovak phonemes were used based on the linguistic knowledge. The task was then to find the "optimal" values for the minimal increase of the log likelihood when splitting a cluster and the minimal occupancy count. The bigger the likelihood and the occupancy count the fewer clusters will be formed which leads to less accurate but more robust models. Thus a right trade off must be found. Original settings in the MASPER procedure are the same as in the HTK book example and these were set to 350 and 100 for the log likelihood and the occupancy count, respectively. In the following, two occupancy counts were tested, the original 100 which seems to be the minimal reasonable option for complex models with 32 mixtures, and a 160 count. For both of them the log likelihoods were tested ranging from 50% more to 70% less compared to the preset value. Tests were done separately for isolated application words that exhibit wider spectrum of used CD phonemes and the digit strings test which has higher perplexity but uses only few phonemes. Results are shown in Fig. 6 for the digit strings test averaged over all tied CD models using 5 different values of the minimal log likelihood for each occupancy count. The same results are shown in Fig. 7 for application words.

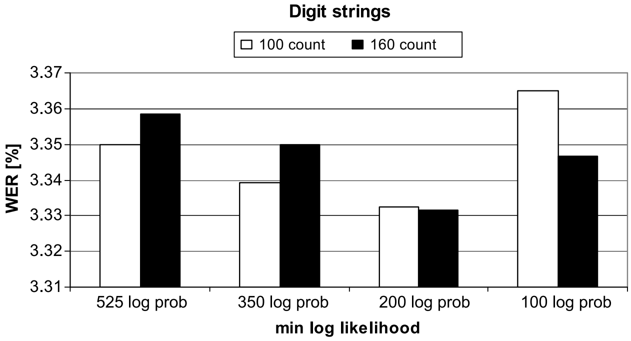

Fig. 6 Effect of the minimal increase of the log likelihood on WER for 100 and 160 occupancy counts in the case of digit strings

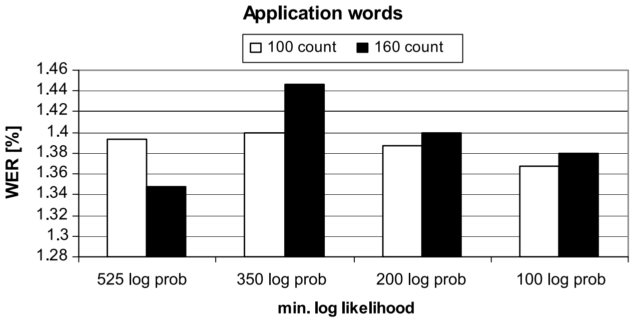

Fig. 7 Effect of the minimal increase of the log likelihood on WER for 100 and 160 occupancy counts in the case of application words

As it can be seen different courses of WER were obtained. For digit strings (Fig. 6) it is just the case of a right trade off, neither too robust nor too accurate. Thus the minimal increase of the log likelihood should be kept somewhere in the middle. Further, for both boundary cases such occupancy count is preferable which eliminates the influence of the "extreme" log likelihood, i.e. for 525 the lower value (100) of the occupancy count is better and for log likelihood equal to 100 the 160 occupancy count shows better results. All these finding are according to the expectations (robust vs. accuracy balance). However, in the application words test (Fig. 7) that contains wider range of CD models and fewer training samples, both extreme cases show slightly better outcomes, i.e. to have either very accurate models or very robust ones (at least in the tested ranges), that is obviously against any assumptions. To get a relevant explanation, particular results from application words test must be look at separately (Fig. 8) instead of having them aggregated.

Figure 8 shows grouped results for different values of the minimal log likelihood as a function of the number of mixtures per model. An obvious trend can be observed that in the case of low complexity models (low number of mixtures) higher values of the likelihood are favorable (fewer divisions occur) whereas for more complex models lower values perform better. This suggests that there is little meaning to have too many different (specialized) models with low complexity as these models lack the modeling abilities. This leads to the construction of more models that have limited discriminative options so these different models are inevitably similar to each other and thus easily mistaken. Therefore it is better to let them be robust and allow the splitting only for very distinct cases that can be properly distinguished even by simple models. The opposite is true for complex models where the higher number of mixtures increases their modeling capacity so even small differences can be described accurately. This would end up in the higher number of distinctive and more precise models. However, the need for a balance between the accuracy and robustness in the case of limited data is obvious when investigating any particular course of WER for a given log likelihood and application words test. As it can be seen form Fig. 8 there is an obvious overtraining phenomenon, as the best results are obtained for middle complex models (from 4 to 16 mixtures). Therefore there are three aspects that are acting together: the complexity of final models (number of mixtures), number of distinctive groups of states, and the existence of limited training data. As there is a greater variety of CD phonemes and less data for application words test the overtraining phenomenon takes place (in the tested ranges). Thus the averaged errors over different models do not provide expected "V" like shape behavior as there are more acting factors. Namely, the overtraining interferes with the finding that is better to have higher number of separate states for more complex models and fewer distinctive states for less complex models.

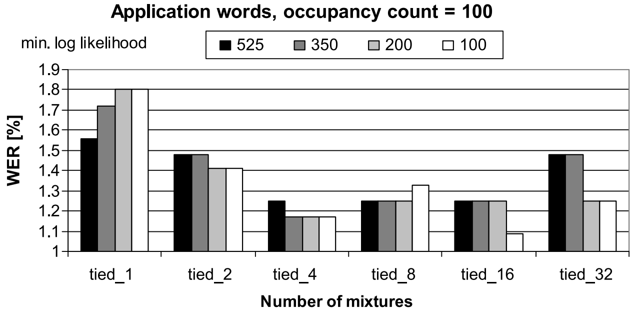

Fig. 8 Impact of the minimal increase of log likelihood on models with different number of mixtures in the case of application words

For digit strings and the tested tying ranges the WER has monotonically declining course as the number of mixtures is increasing regardless of the log likelihood and occupancy counts, Fig. 9.

This is because there were many training examples for relatively limited set of phonemes so even complex models had abundance of training data (there were over 5500 realizations for every single digit) thus any reasonable increase of mixtures caused the drop in error rates regardless of the tested tying options (no overtraining). To conclude the discussion the final optimal settings would depend on the type

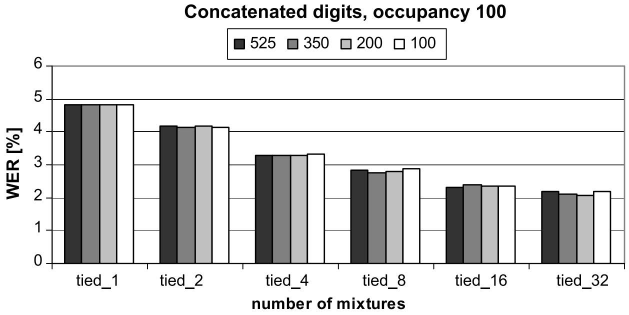

Fig. 9 Impact of the minimal increase of log likelihood on models with different number of mixtures in the case of digit strings

of the application and available database, so it is wise to experiment with different settings having in mind the outlined behavior and relations; however in the most reasonable cases the differences are not substantial. If the final application is a dialog like with finite grammar (it is known what sort of information it is to be recognized in which stage) the designer can make a use of models trained with different setting and of different complexity in each stage of the dialog as the best results are provided by different models.

5 Training database and the testing settings

A vital aspect to succeed in building an accurate and robust speaker independent recognition system is the selection of the proper training database. So far many databases following different employment assumptions and designing goals have been designed and compiled, like: AURORA, TIMIT, SPEECHDAT, SPEECON, etc. Furthermore, the task of recognition is more challenging in adverse environments and requires more steps, additional pre-processing and more sophisticated handling. Since we wanted to demonstrate to the full extend the capabilities, options, modification and pitfalls of the HMM training process, we decided to use the Slovak MOBILDAT database [12] which was recorded over GSM networks and generally provides more adverse environments (wider range of noises, lower SNRs, distortions by compression techniques and short lapses of connections). The concept of MOBILDAT database is based on the widely used structure of the SPEECHDAT database, for which many versions have been built in several languages using fix telephone lines and are regarded as professional databases.

The MOBILDAT-SK database consists of 1100 speakers that are divided into the training set (880) and the testing set (220). Each speaker produced 50 recordings (separate items) in a session with the total duration ranging between 4 to 8 minutes. These items were categorized into the following groups: isolated digit items (I), digit/number strings (B,C), natural numbers (N), money amounts (M), yes/no questions (Q), dates (D), times (T), application keywords (A), word spotting phrase (E), directory names (O), spellings (L), phonetically rich words (W), and phonetically rich sentences (S, Z). Description files were provided for each utterance with an orthographical transcription but no time marks were supplied. Beside the speech, following non- speech events were labeled too: truncated recordings (∼), mispronunciation (∗), unintelligible speech (∗∗), filed pauses (fil), speaker noise (spk), stationary noise (sta), intermitted noise (int), and GSM specific distortion (%). In total there are 15942 different Slovak words, 260287 physical instances of words, and for 1825 words there are more than one pronunciation listed (up to 5 different spellings are supplied). Finally, there are 41739 usable speech recordings in the training portion, containing 51 Slovak phonemes, 10567 different CD phonemes (word internal) and in total there are slightly more than 88 hours of speech.

Accompanying the MASPER training procedure, there are three test scripts working with test sessions defined in the database. These tests are: Small vocabulary isolated phrases (SVIP) contained in I marked recordings, medium vocabulary isolated phrases (MVIP) working with A recordings, and small vocabulary word loop (SVWL) applied to B, C recordings. Tests are however performed only over utterances free of mispronunciation, unintelligible speech and truncated speech. Abovementioned tests are applied to all models that were produced during the training so that a possible overtraining can be detected, and the best models chosen. Both word error rates and sentence error rates are computed, however more common are WER. In the evaluation process the recognized words are aligned with the transcription so that the minimal mistake is achieved using the number of substituted words, deleted and inserted words [25]. Before the evaluation, all marks for non- speech events are removed from the transcription files and in the case of digits two word mappings are performed.

6 Conclusions

Even though there are many new and improved techniques for HMM modelling of speech units and different feature extraction methods, still they are usually restricted to laboratories or specific conditions. Thus most of the practical systems designed for large vocabulary and speaker independent tasks use the "classical" HMM modelling by CDHMM with multiple Gaussian mixtures and tied CD models of phonemes.

In this article the construction of robust and accurate HMM models was presented using one of the most popular system and the training scheme. All the suggested and presented modifications were tested on the professional

MOBILDAT-SK database that poses more adverse environment. Three issues were tackled: feature extraction methods, HMM training schemes, and tying process for CD phonemes.

In the case of extraction methods, PLP an MFCC with their auxiliary features were tested. It was observed that PLP features themselves are better in describing the static speech for recognition purposes. The incorporation of dynamic features is vital for both methods; however their contribution for PLP was more relevant. On the contrary the inclusion of C0 (static feature) and application of the cepstral mean subtraction (constant processing) was much more beneficial in the case of MFCC. Furthermore, the voicing feature was tested as well and proved to be surprisingly beneficial for both methods; the average improvement for PLP was 19.96% and 24.51% for MFCC.

Further a modification to the MASPER training scheme was designed and tested. Its main contribution is to save some useful data from the damaged recordings by constructing BH model for garbled speech. By doing so the number of phonemes' instances increased by more than 9%, triphone instances on average by 8.7%, and 0.6% more triphones was found. This led to more than a 5% improvement for complex models of CD phonemes. More BH models were tested with different training strategies, were the best one uses a 3 state ergodic model trained alone using CI models derived from the "health" recordings and applied only to the training of CD phonemes.

Finally, tying options for CD phoneme models using the decision trees and a linguistic phoneme classification were tested. Within the examined ranges the differences in achieved results were relatively minor, however different behavior was observed for different models and tests. In the case of the existence of many triphones on limited data an overtraining phenomenon was observed for more complex models. In this case it was further observed that for less complex models, fewer distinctive states perform better, whereas for more complex models (more mixtures) higher number of separate states (not tied) is better. This can be viewed as that the less complex models do not have the required modeling capability, thus there is no use to have many distinctive states that can be easily mistaken. However, setting proper options is a tricky task that is database and application specific. For different applications different models and settings perform better. This suggests using several sets of models that can be switched among according to the recognition task (digit items, names, applications words, etc.). It should be noted that some of the presented modifications and settings were successfully used while building Slovak ASR system [22].

Acknowledgement The author hereby declares that the article was funded by the following projects: AV 4/020/07, VEGA–1/0718/09.

References

1. Nouza, J., Zdansky, J., David, P., Cerva, P., Kolorenc, J., & Nejedlova, D. (2005). Fully automated system for Czech spoken broadcast transcription with very large (300K+) lexicon. In *Proceedings of interspeech 2005*, Lisbon, Portugal, September, 2005 (pp. 1681–1684). ISSN 1018-4074.
2. Baum, L., & Eagon, J. (1967). An inequality with applications to statistical estimation for probabilities functions of a Markov process and to models for ecology. *Bulletin of the AMS*, *73*, 360–363.
3. Huang, X., Ariki, Y., & Jack, M. (1990). *Hidden Markov models for speech recognition*. Edinburg University Press.
4. Jiang, H., & Li, X. (2007). A general approximation-optimization approach to large margin estimation of HMMs. In *Robust speech recognition and understanding*. I-Tech education and publishing, Croatia, ISBN 978-3-902613-08-0.
5. Bonafonte, A., Vidal, J., & Nogueiras, A. (1996). Duration modeling with expanded HMM applied to speech recognition. In *Proceedings of ICSLP 96*, Philadelphia, USA (Vol. 2, pp. 1097–1100). ISBN: 0-7803-3555-4.
6. Casar, M., & Fonllosa, J. (2007). Double layer architectures for automatic speech recognition using HMM. In *Robust speech recognition and understanding*. I-Tech education and publishing, Croatia. ISBN 978-3-902613-08-0.
7. Hermasky, H., & Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, *2*(4).
8. Nadeu, C., & Macho, D. (2001). Time and Frequency Filtering of Filter-Bank energies for robust HMM speech recognition. *Speech Communication*, *34*.
9. Cheng, O., Abdulla, W., & Salcic, Z. (2005). Performance evaluation of front-end processing for speech recognition systems. School of Engineering Report No. 621, Electrical and Computer Engineering Department, School of Engineering, The University of Auckland.
10. Haque, S., Togneri, R., & Zaknich, A. (2009). Perceptual features for automatic speech recognition in noisy environments. *Speech Communication*, *51*, 58–75.
11. Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, *11*, 95–103.
12. Darjaa, S., Rusko, M., & Trnka, M. (2006). MobilDat-SK–a mobile telephone extension to the SpeechDat-E SK telephone speech database in Slovak. In *Proceedings of the 11-th international conference speech and computer (SPECOM'2006)*, St. Petersburg, Russia (pp. 449–454).
13. Zgank, A., Kacic, Z., Diehel, F., Vicsi, K., Szaszak, G., Juhar, J., & Lihan, S. (2004). The Cost 278 MASPER initiative—crosslingual speech recognition with large telephone databases. In *Proceedings of language resources and evaluation (LREC)*, Lisbon (pp. 2107–2110).
14. Lindberg, B., Johansen, F., Warakagoda, N., Lehtinen, G., Kacic, Z., Zgang, A., Elenius, K., & Salvi, G. (2000). A noise robust multilingual reference recognizer based on SpeechDat(II). In *Proceedings of ICSLP 2000*, Beijing, China, October 2000.
15. Rabiner, L., & Juan, B. (1993). *Fundamentals of speech recognition*. New Jersey: Prentice Hall. ISBN 0-13-015157-2
16. Hönig, F., Stemmer, G., Hacker, Ch., & Brugnara, F. (2005). Revising perceptual linear prediction (PLP). In *Proceedings of INTERSPEECH*, Lisbon, Portugal, Sept. 2005 (pp. 2997–3000).
17. Lee, K., Hon, H., & Reddy, R. (1990). An overview of the SPHINX speech recognition system. *IEEE Transactions on Acoustics Speech and Signal Processing*, *38*(1).
18. Hermansky, H., Hanson, B. A., & Wakita, H. (1985). *Perceptually based linear predictive analysis of speech*. New York: IEEE.
19. Rabaoui, A., Kadri, H., Lachiri, Z., & Ellouze, N. (2008). Using robust features with multi-class SVMs to classify noisy sounds. In *ISCCSP*, Malta.

20. Cheveigne, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. Journal of the Acoustical Society of America, **111**(4).
21. Kacur, J., & Rozinaj, G. (2009). Adding voicing features into speech recognition based on HMM in Slovak. In *IWSSIP09*, Greece.
22. Juhar, J., Ondas, S., Cizmar, A., Rusko, M., Rozinaj, G., & Jarina, R. (2006). Galaxy/VoiceXML based spoken Slovak dialogue system to access the Internet. In *ECAI 2006 workshop on language-enabled educational technology and development and evaluation of robust spoken dialogue systems*, Riva del Garda, Italy, August 29, 2006 (pp. 34–37).
23. Johansen, F. T., Warakagoda, N., Lindberg, B., et al. (2000). The cost 249 SpeechDat multilingual reference recognizer. In *2nd international conference on language resources and evaluation (LREC-2000)*, Athens, May 2000.
24. Höge, H., Draxler, C., Van den Heuvel, H., Johansen, F. T., Sanders, E., & Tropf, H. S. (1999). SpeechDat multilingual speech databases for teleservices: across the finish line. In *Proc. Europ. conf. speech proc. and techn. (EUROSPEECH)*.
25. Young, S., Evermann, G., & Hain, T. (2002). *The HTK book V.3.2.1*. Cambridge University Engineering Department.
26. Kacur, J., & Ceresna, M. (2007). A modified MASPER training procedure for ASR systems and its performance on Slovak MOBILDAT database. In *IWSSIP07*, Slovenia.

Juraj Kačur born in Bratislava in 1976. Master of Science degree (M.Sc.) obtained in Jan. 2000 and Ph.D. in 2005 at the faculty of electrical engineering and information technology of the Slovak university of technology (FEI STU) Bratislava.

Between Jan. 2000 and Feb. 2001 he was with the Slovak academy of science, department of speech analysis and synthesis where he participated on several projects. Since Feb. 2001 he occupies an assistant professor position at the department of telecommunication at FEI STU Bratislava. The field of his research activities includes: digital speech processing, speech recognition, speech detection, speaker identification, High order statistic, Wavelet transform, ANN and HMM.

Gregor Rozinaj (M'97) received M.Sc. and Ph.D. in telecommunications from Slovak University of Technology, Bratislava, Slovakia in 1981 and 1990, respectively.

He has been a lecturer at the Department of Telecommunications of the Slovak University of Technology since 1981. From 1992–1994 he worked on the research project devoted to speech recognition at Alcatel Research Center in Stuttgart, Germany. From 1994–1996 he was employed as a researcher at the University of Stuttgart, Germany working on a research project for automatic ship control. Since 1997 he has been a Head of the DSP group at the Department of Telecommunications of the Slovak University of Technology, Bratislava. Since 1998 he has been an Associate Professor at the same department. He is an author of 3 US and European patents on digital speech recognition and 1 Czechoslovak patent on fast algorithms for DSP.

Dr. Rozinaj is a member of IEE and IEEE Communication Society.