# Three-Stage Framework for Unsupervised Acoustic Modeling Using Untranscribed Spoken Content

Andrej Zgank

**This paper presents a new framework for integrating untranscribed spoken content into the acoustic training of an automatic speech recognition system. Untranscribed spoken content plays a very important role for under-resourced languages because the production of manually transcribed speech databases still represents a very expensive and time-consuming task. We proposed two new methods as part of the training framework. The first method focuses on combining initial acoustic models using a data-driven metric. The second method proposes an improved acoustic training procedure based on unsupervised transcriptions, in which word endings were modified by broad phonetic classes. The training framework was applied to baseline acoustic models using untranscribed spoken content from parliamentary debates. We include three types of acoustic models in the evaluation: baseline, reference content, and framework content models. The best overall result of 18.02% word error rate was achieved with the third type. This result demonstrates statistically significant improvement over the baseline and reference acoustic models.**

**Keywords: Automatic speech recognition, acoustic modeling, untranscribed spoken content, data-driven metric, imperfect transcriptions.**

## I. Introduction

Automatic speech recognition research has experienced enormous development in the last two decades [1]. The number of speech recognition applications that are being transferred from laboratories into a real-life environment is constantly rising [2]-[5]. Automatic speech recognition still relies on manually transcribed and annotated spoken language resources, whose production, regardless of recent developments, is still very expensive and time-consuming. This is one of the main reasons speech recognition technology is available only for "major" languages, that is, some major western European and Asian languages, out of a total of approximately 6,000 world languages.

One of the possible ways [6], [7] to create spoken language resources for training acoustic models is to collect some type of untranscribed spoken content and apply procedures for unsupervised [8] or lightly supervised training [9]. This can improve quality for under-resourced languages. The main idea [6], [10], [11] behind such methods is to train a baseline speech recognition system using an existing, small, manually transcribed speech database, to recognize acquired untranscribed spoken content, and to apply the recognized transcriptions for unsupervised acoustic training or adaptation.

Wessel and Ney [11] presented an unsupervised training method which used a confidence measure. Lamel and others [6] proposed an unsupervised method with incremental chunks of untranscribed content. They also showed that retranscribing can significantly improve performance. Cincarek and others [7] used selective training to decrease the complexity of unsupervised acoustic model training. Some authors [8], [10] have compared the incremental with the "use-all-data" approach for unsupervised training. These methods process

Manuscript received Mar. 14, 2010; revised May 27, 2010; accepted June 28, 2010.

Andrej Zgank (phone: +386 2 220 7206, email: andrej.zgank@uni-mb.si) is with the Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia.

doi:10.4218/etrij.10.1510.0092

  

untranscribed spoken content on the level of acoustic models [12]. Lecouteux and others [13] presented a method in which unsupervised training was incorporated directly into the speech decoder. Very similar methods are also used for lightly supervised training [9], in which some sort of imperfect transcription [13] is used in each iteration.

In this paper, we propose a new framework for processing acquired untranscribed spoken content. The main focus in the proposed framework is on improving the first few iterations of untranscribed content processing in which the portion of speech recognition errors is higher, as was shown in the preceding analysis. This is directly connected with the complexity (high inflection) of the language that is being recognized. In general, the existing methods presented above use a "homogenous" unsupervised training approach. Our framework introduces three stages with various complexities. This approach models complex languages better and handles the content conditions mismatch. In the first stage, we propose a new method for generating initial content acoustic models. This method produces combined initial acoustic models based on a data-driven phoneme confusion matrix metric. The result of the proposed method has some similarity with the maximum a posteriori (MAP) adaptation procedure, that is, "linear combination." However, it generates acoustic models that are even more general and, as such, more appropriate for initial acoustic model training. In the second stage, we propose a new broad-endings training approach, which reduces the impact of acoustic reductions occurring at the ends of words during acoustic models training. This novel unsupervised training method is important for highly inflectional and agglutinative languages, in which word endings represent a high proportion of speech recognition errors.

Untranscribed spoken content can be obtained from various sources, including radio and TV broadcasts [14], [15], meetings of government bodies, and Internet sites. In our case, we used parliamentary debates as a surce of untranscribed spoken content.

Slovenian is taken as an example of an under-resourced language [16]. With only 2 million native speakers, it is one of the smallest official languages in the European Union. Slovenian belongs to the Slavic language group and is a highly inflectional Indo-European language with relatively free word order, which makes the speech recognition task very complex [17]. There are approximately 300 hours of spoken language resources available for Slovenian [16]. These have both perfect and imperfect transcriptions. However, when language complexity is taken into account [18], this amount of spoken language resources cannot be compared with the amount of spoken language resources available for major world languages such as English, Spanish, and Mandarin.

This paper is organized as follows. Section II presents the proposed framework for incorporating the new content into acoustic model training. A short overview of the spoken language resources involved is given in section III. The undertaken experiments are described in section IV. The speech recognition results obtained and their evaluation are presented in section V, and the conclusion and future prospects are given in section VI.

## II. Framework for Unsupervised Acoustic Model Training

We propose a new framework for using the collected untranscribed spoken content to train the baseline acoustic models of a speech recognition system in this section. The framework consists of three stages,

- Stage 1: speech/non-speech preprocessing of content and generation of combined initial acoustic models,
- Stage 2: unsupervised broad-endings acoustic model training,
- Stage 3: final standard unsupervised acoustic model training.

One training iteration cycle consists of two parts: a speech recognition task and acoustic model training. First, the speech recognition task is run using the acoustic models from the previous iteration, which generates the current version of content transcriptions. Any words whose transcriptions are very likely erroneous are omitted from the training set by using the decoder's confidence measure threshold. Then, the acoustic model training is performed, applying this new version of the transcriptions. The number of iterations in each stage is set in an empirical way using a development set.

In order to compare our framework with the reference unsupervised training method, a second set of content acoustic models was created. The unsupervised procedure proposed by Lamel and others [6] was used as the baseline for training the reference acoustic models. Necessary modifications were made to guarantee the comparability of methods and an equal number of unsupervised acoustic model training iterations. The topology and complexity of the final acoustic models trained with our procedure, and the reference procedure were comparable.

### 1. Stage 1: Speech/Non-speech Preprocessing and Generation of Combined Initial Acoustic Models

Analysis of the untranscribed spoken content showed that the speech signal contained longer portions of silence in some places. One reason for these pauses was the time which elapsed during a change of speakers, and another cause may be connection errors (overloaded streaming server, stream quality

change, and network congestion) when content is collected over the Internet. These silent parts are of no use for additional acoustic model training, but they still significantly contribute to the processing time. Therefore, we decided to exclude them from the training set by applying the speech/non-speech classification.

The speech/non-speech classification was carried out using Gaussian mixture models (GMMs), which were trained on 1/20 of the full baseline training set. The non-speech model represented silence and low-level background noise. The speech model had 128 mixture probability density functions, and the non-speech model had 64 mixture probability density functions. The difference in the number of mixtures between the speech and non-speech models arose because of the varying amount of training material available per class. The speech/non-speech classifier output was post-processed to smooth the condition changes which were shorter than four consecutive frames. This reduced the number of glitches in the classifier's output. The speech segments were then used for unsupervised acoustic model training.

Some extra steps were already involved (see section IV) in the feature extraction frontend to reduce the acoustic condition mismatch between the baseline database and spoken content database. The influence of this resource mismatch is especially noticeable during the first iteration. We present a new method to initialize acoustic models for unsupervised training in order to reduce the disadvantage of combining spoken content language resources.

The basic idea behind this initialization method is that, during a phoneme speech recognition task, similar phonemes are confused with one another more often than dissimilar ones. Similar phonemes should use the same acoustic model to generalize the initial acoustic-phonetic space of diverse spoken language resources. Acoustic models used for the phoneme speech recognition task were trained on the Slovenian Broadcast News (BNSI) baseline speech database whereas the recognition set was the training set of the untranscribed spoken content of the Slovenian Parliament (SloParl). A number of phoneme confusions result from this specific phoneme speech recognition task. The reference target transcriptions are prepared as follows. First, the speech recognition task is run at the word level with the source acoustic models. Then, the word level transcriptions are converted into phoneme level reference transcriptions using the phonetic vocabulary of a speech recognition system. The end result is a phoneme confusion matrix.

For the first iteration of unsupervised acoustic training, the combined initial monophone acoustic models, denoted as target acoustic models, were built as the normalized average sum of the few most influential source monophone acoustic models according to the weights calculated from the phoneme confusion matrix.

The new initial target monophone acoustic model $M_{\text{trg}}$ is represented as

$$M_{\text{trg}} = \sum_{i=1}^{N_{\text{P}}} w_i M_{\text{src}_i}, \tag{1}$$

where $M_{\text{src}_i}$ denotes a particular source monophone acoustic model, and $w_i$ denotes the influence weight. The $i$ denotes the current model in the pool of most influential models. The number of source monophone acoustic models in the pool that are included in the calculation of a target monophone acoustic model is limited by the parameter $N_P$. This parameter can either be set manually or using a metric. In our case, we used the following empirical criterion for $N_P$:

$$N_{\text{P}} = \begin{cases} 3, & w_3 \geq 0.1, \\ 2, & w_3 < 0.1. \end{cases} \tag{2}$$

Continuous density hidden Markov Models (HMMs) were used in the system. The model $M_{\text{src}_i}$ in (1) can be decomposed into four components: means, mixture weights, variances, and transition probabilities. The initial target acoustic model means are defined as

$$\boldsymbol{\mu}_{\text{trg}} = \sum_{i=1}^{N_P} w_i \boldsymbol{\mu}_{\text{src}_i}, \tag{3}$$

where $\boldsymbol{\mu}_{\text{trg}}$ and $\boldsymbol{\mu}_{\text{src}}$ denote mean values of the target and source acoustic models, respectively. The influence weight is represented by $w_i$. The next components are mixture weights of the target acoustic model, which are defined as

$$\boldsymbol{\omega}_{\text{trg}} = \sum_{i=1}^{N_P} w_i \boldsymbol{\omega}_{\text{src}_i}, \tag{4}$$

where $\boldsymbol{\omega}_{\text{trg}}$ and $\boldsymbol{\omega}_{\text{src}}$ denote mixture weight values for target and source acoustic models, respectively. The target acoustic model variances are given by the maximal variance of all source probability density functions (PDFs):

$$\boldsymbol{v}_{\text{trg}} = \max_i (\boldsymbol{v}_{\text{src}_i}), \tag{5}$$

where $\boldsymbol{v}_{\text{trg}}$ and $\boldsymbol{v}_{\text{src}}$ denote variances for target and source acoustic models, respectively. Equation (5) has a generalizing effect, which is advantageous for initial acoustic models. The probabilities in the transition matrix are the fourth element of the initial target acoustic model. They are calculated as

$$\alpha_{\text{trg}} = \sum_{i=1}^{N_P} w_i \alpha_{\text{src}_i}, \tag{6}$$

where $\alpha_{trg}$ and $\alpha_{src}$ denote the transition probability for the target and source acoustic models, respectively.

The final step necessary is to define the weight $w_i$, which gives the similarity between a source and target phoneme according to the phoneme confusion matrix. The weight $w_i$ is defined as

$$w_i = \frac{\text{conf}\left(M_{\text{src}_i}, T_{\text{trg}}\right)}{\sum_{j=1}^{N_p} \text{conf}\left(M_{\text{src}_j}, T_{\text{trg}}\right)}, \qquad 1 \le M_{\text{src}_i} \le N_P, \qquad (7)$$

where $\text{conf}\left(\text{M}_{\text{src}_i}, \text{T}_{\text{trg}}\right)$ denotes the number of confusions between the pool of source acoustic models $M_{src}$ and the current target phoneme as labeled in the transcriptions. The source weights $w_i$ are normalized, and so the sum of all weights for one target acoustic model must meet the following criterion:

$$\sum_{i=1}^{N_P} w_i = 1 . \qquad (8)$$

The fulfillment of this criterion is particularly important because it guarantees that the new probability density functions for the target acoustic model will be built correctly.

## 2. Stage 2: Unsupervised Broad-Endings Acoustic Training

The second stage of unsupervised acoustic training starts with the final version of acoustic models from the first. Word endings are frequently truncated during continuous speech due to reductions [19], [20]. This phenomenon is present in many languages. Its influence can be severe in highly inflectional and agglutinative languages featuring many different word endings [17], [21]. The preceding analysis showed that the first few iterations of speech recognition for unsupervised acoustic training thus exclude otherwise regular sentences as outliers because of misrecognized or reduced word endings. The decision of whether or not a sentence is excluded is made using the decoder's confidence measure value. The proportion of such outliers is influenced by the type of content and language complexity, but it can be as high as 10% of the full set.

We propose a new method of unsupervised broad-endings acoustic model training to reduce this drawback. The reduction phenomenon usually has the greatest impact on the final vowel of a word. The idea behind the method we propose is to use modified vocabulary (and consequently, transcriptions) for speech recognition and unsupervised training, in which the ending phonemes are altered into broad phonetic classes (BPCs). The main characteristic of BPCs is that they occupy a larger acoustic-phonetic space than phonemes. The BPCs involved in our framework were generated manually by an expert. There are several methods available for creating the broad phonetic classes in a data-driven manner [22] in order to exclude the need for specific language expertise. This can be important in languages with a small number of speakers.

Table 1. Example of original and modified vocabulary.

| Original orthographic transcription | Original phonetic transcription |
|---|---|
| *hiša* | x i: S a |
| *hiše* | x i: S E |
| *hiši* | x i: S i |
| *hišo* | x i: S O |
| **Modified orthographic transcription** | **Modified phonetic transcription** |
| *hiša* | x i: S BPC1 |
| *hiše* | x i: S BPC1 |
| *hiši* | x i: S BPC2 |
| *hišo* | x i: S BPC2 |

We used the following five BPCs in the framework:

- BPC1: open and open-mid vowels,
- BPC2: close and close-mid vowels,
- BPC3: sonorants,
- BPC4: voiced non-sonorants,
- BPC5: unvoiced non-sonorants.

An example of the proposed vocabulary is given in Table 1. The nominative, genitive, dative, and accusative cases of the Slovenian noun *hiša* "house" are shown here. The noun *hiša* consists of the stem *hiš-* and the ending *-a*. According to the acoustic and phonetic properties of the suffixed phoneme, the applicable broad phonetic class defined above was used in the modified vocabulary, that is, the phoneme /a/ is an open vowel, and thus PBC1 was used.

An additional set of acoustic models was created for the proposed broad-endings training method for the newly defined BPC1 to BPC5 models. The BPC acoustic models were initialized with the flat-training approach using a global mean and variance. The flat-start approach was used due to the error-prone, automatically-generated transcriptions that were used for initialization. Each initialized BPC acoustic model was then trained using the Baum-Welch procedure from modified transcriptions of the training set for each particular class.

Four iterations of the broad-endings acoustic model training were included in the framework. The amount of content involved was increased from 1/4 to 1/1. Incrementally increasing the training data size is a technique frequently applied in unsupervised training procedures [6], [7], [9] because it helps manage the complexity of task.

### 3. Stage 3: Final Standard Unsupervised Acoustic Model Training

After the broad-endings acoustic models for the second stage were built, several iterations of standard unsupervised acoustic training were run in the third stage. The original phonetic vocabulary was applied in this stage, and so the true phonemes were used for endings instead of broad phonetic classes. The unsupervised acoustic model training method proposed by Lamel and others [6] was used as a baseline for this stage. Instead of first adding new content in chunks [6], the method was modified in order to use only the complete untranscribed content set. This modification was a result of the first two stages, in which the content amount was increased incrementally. The unsupervised training method on the complete set had already yielded good results in [6].

The result of this framework was a set of final acoustic models, which were trained on the speech database created from the untranscribed content.

## III. Spoken Language Resources

The baseline spoken language resource used in these experiments was the BSNI speech database [23]. This database was designed in cooperation with the University of Maribor and the Slovenian national broadcasting company, RTV Slovenia. It consists of two different types of TV news shows. The first type is the evening news, which gives an overview of daily events, and the other type is the late-night news program, which focuses on two to three major topics. The transcriptions were manually annotated and transcribed according to recommendations on building broadcast news spoken language resources, which were produced in connection with the Transcriber tool [24]. This tool was used for transcription and manual segmentation.

There are a total of 36 hours of manually transcribed speech material available for experiments. The size of the training set is 30 hours. The BNSI database has 1,565 different speakers, of which are 1,069 male, and 477 are female. The gender of the remaining 19 speakers was annotated as unknown. This gender bias in the database reflects the actual proportion of male to female speakers on Slovenian news shows.

The secondary spoken language resource that was used to represent untranscribed content was the SloParl database [25]. Imperfect, error-prone transcriptions generated by the Parliamentary Transcription Office were also available as part of this database, but they were not used in our framework. These imperfect transcriptions are available from the Parliamentary Internet Archive website. This type of content is well-suited to generating untranscribed spoken language resources quickly and economically. The sources of such content are quite frequently various governmental sites and media producers that are obliged to make their production available accessible to handicapped persons.

The SloParl database consists of 100 hours of spoken debates. Although the discourse type is political debates, the SloParl database covers fairly broad topics because a large variety of national issues are covered in the set. Therefore this presented no limitation to constructing the experimental setup. The amount of speakers in this database is smaller in comparison to the BNSI database as it has only 255 different speakers. Acoustically, the spoken material is comparable to the BNSI database, although some level of reverberation is sometimes present in the recordings.

Special requirements had to be taken into account for the evaluation procedure due to the framework's particular properties:

- The focus had to be on improving the acoustic modeling.
- The impact of using diverse content for acoustic modeling such as transfer of databases and channel conditions had to be minimized.
- The impact of acoustic-phonetic characteristics of words had to be taken into account.

We decided to use the third speech database for evaluation to meet these requirements. The Slovenian PoliDat speech database is used to develop voice-driven telephone services. The difference in speaking style between the two speech content databases is not as large as would be expected because the majority of speakers in parliament prepare debates in advance or even read their statements. The PoliDat database is constructed according to the specifications for the SpeechDat databases [26]. Recordings of 1,200 speakers were made using stationary, mobile, and VoIP phones. A set with phonetically balanced, isolated words was taken from the PoliDat database for evaluation. This test set fulfilled all the requirements given above. Each of the 200 test speakers pronounced four different phonetically-balanced words. This allowed us to avoid using a statistical language model, which could have masked the differences in acoustic modeling. The speech recognizer's vocabulary has 1,491 different words, enabling full coverage of the applied test set. The speakers in the PoliDat database were selected so the demographic characteristics of gender, age, and accent of the complete population were fairly represented.

The BNSI and SloParl databases were recorded using a wideband channel and the PoliDat database used was produced with a narrowband channel. To neutralize this discrepancy, special steps were needed during the acoustic preprocessing phase (see section IV). Detailed statistics for all three spoken

Table 2. Spoken language resources.

| | BNSI | SloParl | PoliDat |
|---|---|---|---|
| Length (h) | 36 | 100 | 21 |
| No. of speakers | 1,565 | 255 | 1,200 |
| No. of words | 268,000 | 655,000 | – |
| Vocabulary | 37,000 | 37,000 | – |

language resources are given in Table 2.

The data in Table 2 clearly shows how important the untranscribed spoken content resources are in speech recognition. The SloParl database has the largest amount of spoken material available, although its creation was simpler. If more untranscribed material is needed, this production process could be repeated with the same content source or a new one.

## IV. Experiments

The experiments were based on a speech recognition system that uses continuous-density HMMs with Gaussian PDFs. The HTK toolkit was used for HMM acoustic modeling.

We converted the BNSI and SloParl databases to an 8 kHz narrowband version during the acoustic preprocessing to match the PoliDat database conditions. After the down-sampling an additional band-pass filter, which simulated telephone channel conditions, was constructed using Matlab and applied to all down-sampled recordings. This assured identical acoustic conditions for all three spoken language resources.

The frontend applied for constructing a speech recognition system was based on mel-frequency cepstral coefficients (MFCC) and energy (12 MFCC + 1 E, delta, delta-delta). The feature vector size was 39. The cepstral mean normalization was added to the feature extraction procedure to further reduce the differences in acoustic channels between all speech databases [27].

The baseline speech recognition system was developed using the BNSI speech database before the proposed framework was applied. The manually segmented speech material was used for baseline training. This was necessary to exclude the impact of potential errors that could occur during the automatic segmentation procedure on our baseline acoustic models. The baseline acoustic models were generated in three steps. First, the context-independent acoustic models with one mixture of Gaussian PDF were built. Their task was to improve the original BNSI transcriptions using the forced realignment procedure. Utterances that were too divergent with the first acoustic models were excluded as outliers. The amount of such utterances was approximately 0.35% on the first run.

The second step was similar to the first one. Acoustic models were once again trained from scratch using the refined transcriptions generated in the first step. This time, context-independent acoustic models with a mixture of Gaussian PDFs were trained before the forced alignment procedure was applied once again. The amount of outliers decreased to 0.21%, which reflects the level of improvement.

In the final step of building the baseline system, triphones, the context-dependent acoustic models, were trained. The number of free triphone parameters was reduced using the phonetic decision-tree based clustering method. The decision trees were induced with the broad phonetic classes created from a data-driven approach based on the phoneme confusion matrix [22]. Prior results showed that this decision-tree based clustering approach improves the speech recognition performance. The number of Gaussian PDFs per state was incrementally increased to 16 when the clustering was finished. These acoustic models were used as the final baseline set.

This baseline speech recognition system functioned as the starting point for our framework, in which additional training data based on the content collected was added. The acquired speech material was first segmented using the GMMs to exclude the non-speech parts. The initial acoustic models were created by combining BNSI baseline acoustic models using the phoneme confusion matrix metric during the first stage. The first-stage initial acoustic models were refined during the second stage with the unsupervised broad-endings acoustic training. Four unsupervised broad-endings training iterations applying various amount of content were carried out.

After an additional four standard iterations of the unsupervised acoustic training procedure, the final content acoustic models included in the evaluation of this framework were created.

For the evaluation procedure, a reference content-based system was trained in unsupervised mode. This system had the same number of iterations as were carried out in the framework system.

The full experimental setup used a large amount of spoken material, which was involved in several iterations. We used a group of multi-core servers as a hardware platform in order to be able to manage this task in a reasonable time.

## V. Results

The first evaluation step was devoted to analyzing the baseline set of acoustic models trained on the BNSI speech database. All speech recognition results are given as word error rate (WER). The phonetically-balanced isolated words from the PoliDat database were used for evaluation. The WER for the baseline acoustic models is given in Table 3.

Table 3. Speech recognition results on PoliDat database with BNSI baseline system.

| Acoustic models | WER (%) |
|---|---|
| BNSI baseline | 23.63 |

Table 4. Speech recognition results on PoliDat database with reference content acoustic models.

| Acoustic model version | No. of iteration | Content amount | WER (%) |
|---|---|---|---|
| Cont-ref-1 | 1 | 1/4 | 20.69 |
| Cont-ref-2 | 2 | 1/3 | 21.09 |
| Cont-ref-3 | 3 | 1/2 | 19.89 |
| Cont-ref-4 | 4 | 1/1 | 19.63 |
| Cont-ref-7 | 7 | 1/1 | 19.50 |
| Cont-ref-8 | 8 | 1/1 | 19.50 |

The overall baseline WER of 23.63% is comparable with results achieved on other systems of similar complexity [28], [29]. The main factors contributing to the relatively high WER are the vocabulary size, database type (combining two spoken language resources), and complexity (phonetically-balanced words) of the test scenario. A speech recognition system of similar complexity in which acoustic models were trained solely on the PoliDat speech database achieved a WER of 15.62%. The difference between the BNSI baseline system and the solely PoliDat system mainly results from the use of various speech databases and all the issues connected with this, that is speaking style, speaker demographics, and topic. It can be anticipated that the difference in WER will be reduced to some extent with the use of a content speech database.

The second evaluation step centered on the reference content acoustic models that were designed for reasons of comparison, that is, without using the new methods proposed here. Various versions of acoustic models in the second evaluation step differ in two aspects: first, in the amount of content material that was used for training and, second, the number of training iterations. The results of the second evaluation step are presented in Table 4.

The speech recognition results for the reference content acoustic models generated with the standard unsupervised acoustic model training method show general improvement over the BNSI baseline, although the decrease in WER is not homogeneous. The first version of the reference content acoustic model, Cont-ref-1, in which only one iteration of acoustic model training on one-quarter of the content material was applied, achieved a WER of 20.69%. The best overall result in the second evaluation step was 19.50%. This was achieved with the Cont-ref-8 version (eighth iteration, full content set) of the acoustic model. The improvement was an absolute 4.13%. The same result was also achieved with the Cont-ref-7 acoustic models, which indicates saturation of the acoustic training procedure. A special case was the Cont-ref-2 set of acoustic models, in which a small degradation of performance occurred in comparison with the models from the previous iteration. In this case, the WER increased from 20.69% to 21.09%. Detailed analysis of the training framework showed that the possible cause could be the additional amount of content material, in which larger proportions of long utterances were observed. This type of utterance can be problematic in unsupervised training because the errors occurring in long utterances accumulate and degrade the training procedure quality.

Table 5. Speech recognition comparison on PoliDat database for the stage 1 acoustic models.

| Acoustic model version | Content amount | WER (%) |
|---|---|---|
| BNSI baseline retrain | 1/4 | 22.70 |
| Stage 1 | 1/4 | 22.16 |

The proposed framework was evaluated in the last step. First, the method for combining initial acoustic models proposed in stage 1 was evaluated. The stage 1 acoustic models were partly compared with the BNSI baseline as shown in Table 3 and partly with the BNSI baseline retrained on the content training set (see Table 5). One iteration of Baum-Welch retraining was used as an adaptation to reduce the database condition mismatch.

The retrained BNSI baseline acoustic models improved the results for the BNSI baseline acoustic models from a WER of 23.63% to 22.70%. The stage 1 acoustic models yielded a WER of 22.16%. These results show that both approaches for reducing condition mismatch improved the speech recognition results, with slightly better improvements from the stage 1 acoustic models.

Second, the entire proposed framework was evaluated. Various versions of acoustic models differed in the amount of content material and the number of training iterations applied. In addition, the combinations of initial acoustic models and broad-endings training were also included. The results of the third evaluation step are presented in Table 6.

The speech recognition results for the framework content acoustic models show improvement in comparison to the reference content acoustic models. The first iteration Cont-frmw-1 acoustic models achieved a WER of 19.89%. The difference between the framework and reference acoustic

Table 6. Speech recognition results on PoliDat database with framework content acoustic models.

| Acoustic model version | No. of iteration | Content amount | WER (%) |
|---|---|---|---|
| Cont-frmw-1 | 1 (ciAM+b_end) | 1/4 | 19.89 |
| Cont-frmw-2 | 2 (b_end) | 1/3 | 19.89 |
| Cont-frmw-3 | 3 (b_end) | 1/2 | 19.50 |
| Cont-frmw-4 | 4 (b_end) | 1/1 | 18.96 |
| Cont-frmw-7 | 7 | 1/1 | 18.69 |
| Cont-frmw-8 | 8 | 1/1 | 18.02 |

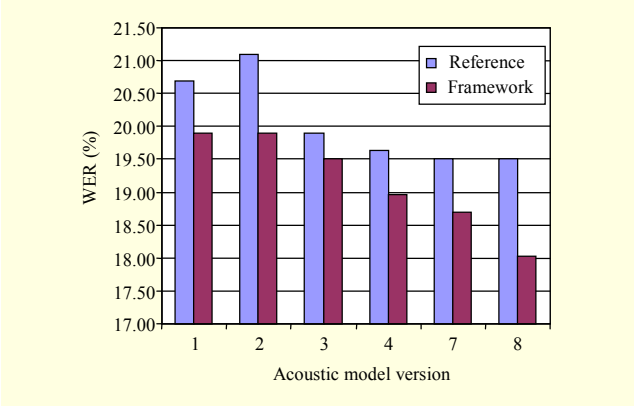


Fig. 1. WER for various versions of reference and framework acoustic models.

models is in the use of combined initial acoustic models and broad-endings approach for training the framework models. In the second iteration, which is one-third of the content, the WER remained the same. No degradation such as that occurring in the reference content acoustic models was seen. This result indicates that the broad-endings training approach can, to some extent, compensate for the problem of accumulating errors in long utterances. In the last iteration of the second framework stage (Cont-frmw-4) the WER was 18.96%. Comparison with the reference content acoustic models shows improvement, although some of the improvement achieved during the first three iterations was lost. This indicates that the use of the broad-endings training procedure is very reasonable in cases in which only a few training iterations are possible due to system limitations.

In the third iteration of the third stage (Cont-frmw-7) the standard unsupervised acoustic model training was used. A WER of 18.69% was obtained. Compared to the reference content acoustic models, these acoustic models preserved the improvements achieved during the first two stages and also gained some room for additional improvements, which was lacking in the second half of the reference version (see Table 4). This probably results from the extended coverage of the acoustic-phonetic space that is one of the properties of the combined initial acoustic models.

The final framework content acoustic model was Cont-frmw-8. It achieved a WER of 18.02%, which is 1.48% absolute improvement over the reference content acoustic models. The framework content acoustic models outperformed the BNSI baseline acoustic models by 5.61%, which clearly shows the advantage of using additional spoken language resources automatically generated from untranscribed spoken content. The best overall result is still 2.40% less accurate than that of the solely PoliDat system. Because the third stage of the framework system showed no training procedure saturation, it can be assumed that an additional amount of untranscribed spoken content could further reduce this gap. Figure 1 shows the performance improvement graph for using untranscribed content.

## VI. Conclusion

This paper presented a new framework for additional training of speech recognizer acoustic models using untranscribed spoken content. Two new methods were proposed in the scope of this framework: a combination of initial acoustic models based on a phoneme confusion matrix metric and unsupervised broad-endings acoustic model training. The evaluation of speech recognition results showed a significant decrease in WER compared to the reference content acoustic models. Another advantage of this framework is that it tends to produce acoustic models that can be used for further additional training with new untranscribed spoken content. The main disadvantage of this framework is that it adds two new methods to the setup. This results in increased complexity and the need for phonetic expertise.

Future work will be focused on excluding the need for expert knowledge by introducing a data-driven metric and on further improving combined acoustic models, whereby the primary focus will be on the robustness of modeling the acoustic-phonetic space.

## References

[1] C.H. Lee, "On Automatic Speech Recognition at the Dawn of the 21st Century," *IEICE Trans. Inf. Syst.,* vol. E86-D, no. 3, Mar. 2003, pp. 377-396.

[2] H.Y. Jung, B.O. Kang, and Y. Lee, "Model Adaptation Using Discriminative Noise Adaptive Training Approach for New Environments," *ETRI J.*, vol. 30, no. 6, Dec. 2008, pp. 865-867.

[3] J. Na, W. Choi, and D. Lee, "Design and Implementation of a

Multimodal Input Device Using a Web Camera," *ETRI J.*, vol. 30, no. 4, Aug. 2008, pp. 621-623.

[4] S. Kim, M. Ji, and H. Kim, "Noise-Robust Speaker Recognition Using Subband Likelihoods and Reliable-Feature Selection," *ETRI J.*, vol. 30, no.1, Feb. 2008, pp. 89-100.

[5] T. Cincarek et al., "Development, Long-Term Operation and Portability of a Real-Environment Speech-Oriented Guidance System," *IEICE Trans. Inf. Syst.,* vol. E91-D, no. 3, 2008, pp. 576-587.

[6] L. Lamel, J.L. Gauvain, and G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training," *Computer Speech & Language,* vol. 16, no. 1, 2002, pp. 115-129.

[7] T. Cincarek et al., "Cost Reduction of Acoustic Modeling for Real-Environment Applications Using Unsupervised and Selective Training," *IEICE Trans. Inf. Syst.,* vol. E91-D, no. 3, 2008, pp. 499-507.

[8] S. Novotney, R. Schwartz, and J. Ma, "Unsupervised Acoustic and Language Model Training with Small Amounts of Labelled Data," *Proc. 2009 IEEE Int. Conf. Acoustics, Speech Signal Process.,* Apr. 19-24, 2009, pp. 4297-4300.

[9] B. Chen, J.W. Kuo, and W.H. Tsai, "Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription," *ICASSP*, 2004, pp. 777-780.

[10] J. Ma and R. Schwartz, "Unsupervised Versus Supervised Training of Acoustic Models," *INTERSPEECH,* 2008, pp. 2374-2377.

[11] F. Wessel and H. Ney, "Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition," *ASRU Workshop*, 2001, pp. 307-310.

[12] P.J. Jang and A.G. Hauptmann, "Improving Acoustic Models with Captioned Multimedia Speech," *IEEE Int. Conf. Multimedia Computing Syst.*, Florence, Italy, 1999, pp. 767-771.

[13] B. Lecouteux et al., "Imperfect Transcript Driven Speech Recognition," *Interspeech-ICSLP,* Pittsburgh, PA, 2006, pp. 1626-1629.

[14] A. Lambourne et al., "Speech-Based Real-Time Subtitling Services," *Int. J. Speech Technol.*, vol. 7, no. 4, 2004, pp. 269-279.

[15] J. Brousseau et al., "Automatic Closed-Caption of Live TV Broadcast News in French," *Proc. Eurospeech*, Geneva, Switzerland, Sept. 2003, pp. 1245-1248.

[16] Z. Kačič, "Importance of Merging the Research Potentials for Surpassing the Language Barriers in the Frame of Next Generation Speech Technologies," *Proc. Inf. Soc. Multi-Conf.*, Ljubljana, Slovenia, Oct. 2002, pp. 111-115.

[17] M.S. Maučec, Z. Kačič, and B. Horvat, "Modelling Highly Inflected Languages," *Inf. Sciences*, vol. 166, no. 1, Oct. 2004, pp. 249-269.

[18] A. Žgank, Z. Kačič, and B. Horvat, "Large Vocabulary Continuous Speech Recognizer for Slovenian Language," *Lecture Notes Computer Science*, Springer Verlag, 2001, pp. 242-248.

[19] S. Furui et al., "Analysis and Recognition of Spontaneous Speech Using Corpus of Spontaneous Japanese," *Speech Commun.*, vol. 47, no. 1-2, Sept. 2005, pp. 208-219.

[20] F. Stouten et al., "Coping with Disfluencies in Spontaneous Speech Recognition: Acoustic Detection and Linguistic Context Manipulation," *Speech Commun.*, vol. 48, no. 11, 2006, pp. 1590-1606.

[21] K.N. Lee and M. Chung, "Morpheme-Based Modeling of Pronunciation Variation for Large Vocabulary Continuous Speech Recognition in Korean," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 7, July 2007, pp. 1063-1072.

[22] A. Žgank, B. Horvat, and Z. Kačič, "Data-Driven Generation of Phonetic Broad Classes Based on Phoneme Confusion Matrix Similarity," *Speech Commun.*, vol. 47, no. 3, 2005, pp. 379-393.

[23] A. Žgank et al., "BNSI Slovenian Broadcast News Database: Speech and Text Corpus," *9th European Conf. Speech Commun. Technol.*, Interspeech Lisboa, Lisbon, Portugal, Sept. 4-8, 2005.

[24] C. Barras et al., "Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production," *Speech Commun.*, vol. 33, no.1-2, 2001, pp. 5-22.

[25] A. Žgank et al, "SloParl: Slovenian Parliamentary Speech and Text Corpus for Large Vocabulary Continuous Speech Recognition," *Proc. INTERSPEECH*, *ICSLP*, Pittsburgh, PA, 2006, pp. 197-200.

[26] H. Heuvel et al., "Annotation in the SpeechDat Projects," *Int. J. Speech Technology*, vol. 4, no. 2, 2001, pp. 127-143.

[27] D. Kim and D. Yook, "A Closed-Form Solution of Linear Spectral Transformation for Robust Speech Recognition," *ETRI J.*, vol. 31, no. 4, Aug. 2009, pp. 454-456.

[28] A. Žgank et al., "The COST 278 MASPER Initiative: Crosslingual Speech Recognition with Large Telephone Databases," *Proc. LREC*, Lisbon, Portugal, May 2004, pp. 2107-2110.

[29] F.T. Johansen et al., "The COST 249 SpeechDat Multilingual Reference Recogniser," *Proc. LREC,* Athens, Greece, May 2000, pp. 1351-1355.

**Andrej Zgank** received his PhD from the Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia, in 2003. He was a senior researcher at the University of Maribor from 2003 to 2008. He is now an assistant professor at the University of Maribor in the field of telecommunications. His research interests are in the areas of multilingual and crosslingual speech recognition, acoustic modeling for large vocabulary continuous speech recognition, and the design of advanced telecommunication services and applications. He is a member of IEEE, IEICE, ISCA, and SDJT.