ABSTRACT

The paper addresses the problem of designing a language independent phonetic inventory for the speech recognisers with multilingual vocabulary. A new clustering algorithm for the definition of multilingual set of triphones is proposed. The clustering algorithm bases on a definition of a distance measure for triphones defined as a weighted sum of explicit estimates of the context similarity on a monophone level. The monophone similarity estimation method based on the algorithm of Houtgast. The clustering algorithm is integrated in a multilingual speech recognition system based on HTK V2.1.1. The ongoing experiments are based on the SpeechDat II databases. So far, experiments included the Slovenian, Spanish and German 1000 FDB SpeechDat (II) database. Current results are very promising. The use of clustering algorithm resulted in a significant reduction of the number of triphones at acceptable level of word and language identification accuracy degradation.

1. INTRODUCTION

The development of speech technology in the last few years raised an interest in the research of the multilingual speech recognition. In order to reduce the complexity of a multilingual recogniser and to reduce the cost of a cross-language transfer of speech technology, the development of methods for the definition of the multilingual phonetic inventories is of increasing concern.

The definition of the multilingual phonetic inventories by exploiting similarities among sounds of different languages is a promising approach. First attempt was reported in [1]. Here the multilingual phonetic inventory, consisting of language-dependent and language-independent speech units, was defined using the data-driven clustering technique. Other attempts based on different distance measures and clustering techniques also followed [2,3,4,5], however, all the work so far was focused on the context independent phoneme modelling (monophones). These experiments have shown that the transition from language dependent monophone set to multilingual inventory of monophones may result in a degradation of recognition accuracy due to the lack of acoustic resolution of the multilingual phoneme set.

The transition from the context independent to context dependent phoneme modelling seems inevitable in order to improve the performance of multilingual speech recognition systems, i.e. the speech recognisers with multilingual vocabulary. The development of a method for the definition of the multilingual set of context dependent phoneme models requires the definition of new clustering criteria.

In this paper, a clustering algorithm for the definition of multilingual set of context dependent phoneme models (triphones) is proposed. The clustering algorithm bases on a distance measure for triphones defined as the combination of explicit estimation of the similarity of the phonemes of left and right contexts and the central phonemes.

2. TRIPHONE DISTANCE MEASURE

The crucial problem concerning the use of triphone modelling is large number of triphone models, which requires large amounts of training data. Since the amount of training data is usually limited many of the triphone speech units are rarely or even never seen during the training. For...
this reason the direct implementation of the distance measures that were defined for the monophones, such as \([1, 2, 3, 4]\) is not appropriate for the definition of multilingual set of triphones.

Our definition of the distance measure for triphones bases on the fact that the triphone is "a monophone in a certain context". Therefore, the similarity of two triphones can be estimated indirectly - by explicitly estimating the similarity of both central phonemes, both left-context phonemes and both right-context phonemes. The similarity of two triphones \(l_1-c_1+r_1\) and \(l_2-c_2+r_2\) \((l, c \text{ and } r \text{ denote the left context - phoneme, right context - phoneme and the central phoneme, respectively})\) was therefore defined as:

\[
S(l_1-c_1+r_1, l_2-c_2+r_2) = L s(l_1, l_2) + C s(c_1, c_2) + R s(r_1, r_2)
\]

where \(s\) denotes the similarity of two phonemes, \(L, C, R\) are the weights for setting the influence of each phoneme - level similarity estimates, and \(S(l_1-c_1+r_1, l_2-c_2+r_2)\) is the resulting similarity of both triphones.

Such definition of distance measure for triphones can be based on any type of phoneme-distance measure \((s\) in Equation 1\). In our case, the phone-distance measure was defined as suggested in [1]:

\[
s(f_i, f_j) = \frac{1}{2} \sum_{k=1}^{N} \left[ c(f_i, f_j) + c(f_j, f_i) - c(f_i, f_i) - c(f_j, f_j) \right]
\]

\[1 \leq i, j \leq N, \ i \neq j \quad (2)\]

where \(s(f_i, f_j)\) denotes the similarity between phonemes \(f_i\) and \(f_j\), \(N\) is the number of phonemes, \(c(f_i, f_j)\) is the number of confusions between phonemes \(i\) and phone \(j\).

Described definition of distance measure for triphones has two major advantages. First it offers an accurate estimation of a triphone similarity (similarity of triphones is likely to be higher in a matching context and vice-versa). Next, such definition can provide a reliable estimation of similarity between triphones even in case of "rare" or "unseen" triphones.

### 3. Clustering Algorithm

Having defined the distance measure for the triphones, the clustering algorithm for automatic identification of the triphones that are similar enough to be equated across the languages was defined. A group of triphones is equated if an average distance among all triphones from the group is less than a predefined threshold \(T\). Average distance among \(M\) triphones was defined as:

\[
S(\varphi_1, \varphi_2, ..., \varphi_M) = \frac{\sum_{k=1}^{M} \sum_{i=1}^{M} S(\varphi_i, \varphi_k)}{\sum_{k=1}^{M} k}
\]

\(\varphi_k, \varphi_j \in (\varphi_1, \varphi_2, ..., \varphi_M), k \neq 1\) (3)

where \(\varphi_k\) denotes the triphone \(l_2-c_1+r_1\), \(\varphi_2, \varphi_3, ..., \varphi_M\) is the group of triphones, \(S(\varphi_1, \varphi_2, ..., \varphi_M)\) is the average distance among all triphones from the group \((\varphi_1, \varphi_2, ..., \varphi_M)\). To find all groups of triphones that complies with the condition from the Equation (3), the following 2-stage search algorithm was applied.

In the first stage, a list of most similar phonemes (poly-phonemes) was defined using the method described in [1]. A partial list of poly-phonemes covering all three languages is given in Table 1.

<table>
<thead>
<tr>
<th>n</th>
<th>Slovene</th>
<th>German</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>2</td>
<td>O</td>
<td>O</td>
<td>o</td>
</tr>
<tr>
<td>3</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>4</td>
<td>l</td>
<td>l</td>
<td>l</td>
</tr>
<tr>
<td>5</td>
<td>t</td>
<td>t</td>
<td>t</td>
</tr>
<tr>
<td>6</td>
<td>m</td>
<td>m</td>
<td>m</td>
</tr>
</tbody>
</table>

Table 1. A partial list of poly-phonemes for the Slovene, German and Spanish language.

In the second stage, the groups of triphones to be equated were identified. The search for these groups was limited to the classes of triphones consisting of triphones with the phonemes of the same poly-phoneme as the central phoneme. For example, the search for the similar triphones was first started among the triphones of all three languages with either Slovenian phoneme a, German phoneme a or Spanish phoneme a as the central phoneme. Next, the search for the groups of similar triphones continued among the triphones with either Slovenian phoneme O, German phoneme O or Spanish phoneme o as the central phoneme, etc. Such limitation of search has proven to significantly improve the convergence of the algorithm for the identification of the groups of similar triphones due to the large number of triphones.

This clustering algorithm outputs the list of triphones that are similar enough to be equated...
across the languages. The unlisted triphones remain language specific. The degree of equated triphones can be adjusted by the threshold $T$. The value of $T$ was derived experimentally (values are given with the experimental results).

### 4. BASELINE RECOGNISER

The speech recognition system was based on HTK V2.1.1 with modified frontend module for enhancing the speech recognition robustness. The acoustic feature vector produced by the frontend module consisted of 24 mel-scaled cepstral, 12 $\Delta$ - cepstral, 12 $\Delta\Delta$ - cepstral, high pass filtered energy, $\Delta$ - energy and $\Delta\Delta$ - energy coefficients. This feature vector was processed using the algorithms for maximum likelihood channel adaptation [8] and linear discriminant analysis [8].

Such frontend module was chosen due to the results of previous tests on connected digits recognition task with 99 speakers of the Slovene speech database SNABI and tests on isolated digits recognition task with the databases SNABI and Voice-Mail (German).

The baseline speech recognition system consisted of three language specific recognisers (Slovene, German and Spanish) operating in parallel. The 3-state left-right topology was selected. The recognizer was initially built with 1 Gaussian mixture component per state. All together 24173 triphone models were defined (Sl.-7146, Ge.-12279, Sp.-4748). Parameter tying using the tree-based clustering algorithm (as implemented in the HTK) reduced the number of triphone models to 13074 (Sl.-3517, Ge.-6517, Sp.-3040). At the end the number of Gaussian mixture components per state was augmented to 32.

In the multilingual experiments, the three language specific recognisers operated in parallel using either three language specific model sets or one multilingual set of triphones where many of language specific triphones are tied and used by all three recognisers.

### 5. SPEECH DATABASES

The experiments were carried out using the speech databases produced in the framework of the SpeechDat II project [7]. These databases provide a realistic basis for developing voice driven teleservices and multilingual systems. The following SpeechDat databases were used:

- Slovenian 1000 FDB SpeechDat(II) [6],
- German 1000 FDB SpeechDat(II),
- Spanish 1000 FDB SpeechDat(II).

In all cases, the corpuses contained utterances of 1000 speakers. 800 speakers were used for the training and the remaining 200 speakers were used for the testing of the system. In all experiments the train and test sets were defined as recommended in SpeechDat II project specification.

Only 80 - 95 % of all utterances were useful for the experiments. Remaining utterances were skipped due to the following reasons:
- unusual pronunciation of digits,
- incomplete utterances (speech was cut off at the beginning or end of the utterance),
- unexpected utterances (background noise, comments, ... ).

The system was trained using all corpuses of the train set, while for the testing the corpuses W1-W4 of all three databases, containing phonetically reach words, were used (total of 2252 utterances containing 1960 different words).

### 6. EXPERIMENTAL RESULTS

The experimentation was started recently. Therefore, the results of initial set of experiments are available at the moment.

The baseline recognizer performance is given in the Table 2. The word accuracy (WA) and the language identification rate (LI) are listed for different recognizer architectures. In the first line of Table 2 the performance of the recognizer with 1 Gaussian mixture component per state is given. The use of tree-based clustering algorithm significantly reduced the total number of the triphone models ($N$) from 24173 to 13074 and improved the word accuracy and language identification rates to 71.99% and 91.6%, respectively. Augmenting the number of Gaussian mixture components to 32 additionally improved the word accuracy and language identification rates.

<table>
<thead>
<tr>
<th>$n$ Gauss.</th>
<th>$N$</th>
<th>WA</th>
<th>LI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24173</td>
<td>60.75%</td>
<td>74.5%</td>
</tr>
<tr>
<td>1</td>
<td>13074</td>
<td>71.99%</td>
<td>91.6%</td>
</tr>
<tr>
<td>32</td>
<td>13074</td>
<td>91.52%</td>
<td>93.1%</td>
</tr>
</tbody>
</table>

Table 2. The baseline recogniser performance.
Experiments with multilingual set of triphones were carried out for the recogniser with 13074 models and 1 Gaussian mixture component per state. The clustering algorithm was started at weights $L=1, C=0, R=1$ (see Equation 1) and at different threshold values producing the multilingual triphone sets of different sizes. The performance of the recognizer using various multilingual triphone sets is given in the Table 3.

Besides the word accuracy and the language identification rate, the global compression rate (GCR) was also followed. The global compression rate (GCR) was defined as:

$$GCR = \sum_{i=1}^{N} \frac{M_i}{T_i} c_i$$  \hspace{1cm} (4)

where $N$ is the number of languages, $T_i$ is the number of trainable models in language $i$, $M_i$ is the number of merged models in language $i$ and $c_i$ is the ratio between the number of trainable models in language $i$ and the number of trainable models in $N$ languages.

Table 3. Performance of the recogniser using various multilingual sets of triphones.

<table>
<thead>
<tr>
<th>$T$</th>
<th>$N$</th>
<th>WA</th>
<th>LI</th>
<th>GCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>5928</td>
<td>61.6%</td>
<td>79.1%</td>
<td>35.9%</td>
</tr>
<tr>
<td>40</td>
<td>6245</td>
<td>64.7%</td>
<td>87.4%</td>
<td>47.8%</td>
</tr>
<tr>
<td>60</td>
<td>7602</td>
<td>65.9%</td>
<td>88.2%</td>
<td>58.2%</td>
</tr>
<tr>
<td>80</td>
<td>8070</td>
<td>67.6%</td>
<td>89.9%</td>
<td>61.7%</td>
</tr>
<tr>
<td>100</td>
<td>8565</td>
<td>69.1%</td>
<td>90.6%</td>
<td>64.7%</td>
</tr>
<tr>
<td>120</td>
<td>9965</td>
<td>70.8%</td>
<td>91.2%</td>
<td>78.7%</td>
</tr>
</tbody>
</table>

The clustering algorithm bases on a definition of distance measure for triphones defined as a weighted sum of explicit estimates of the context similarity on a monophone level. In this case the monophone distance estimation method was based on the algorithm of Houtgast. In future, other methods of monophone distance estimation will be also considered.

In future, the number of SpeechDat databases will be increased in order to provide more reliable assessment of the clustering algorithm efficiency.

### 7. CONCLUSION AND FUTURE WORK

At the moment, the described clustering algorithm is in the evaluation process. Initial experiments have shown that the use of clustering algorithm for the definition of multilingual set of triphones can significantly reduce the number of triphone models. A minor decrease of WA and LI rates was also noticed however overall impression of the clustering algorithm performance seems positive.

In the experiments described so far, only the threshold influence has been investigated. The weights $L, C, R$ from Equation 1 were kept at values 1, 0, 1, respectively. The next set of experiments will also include different values of weights (1, 0.5, 1 or 1, 1, 1, etc.).

When influence of the weights will be determined the final set of experiments with the 32 Gaussian mixture component per state will be carried out.

The clustering algorithm bases on a definition of distance measure for triphones defined as a weighted sum of explicit estimates of the context similarity on a monophone level. In this case the monophone distance estimation method was based on the algorithm of Houtgast. In future, other methods of monophone distance estimation will be also considered.

In future, the number of SpeechDat databases will be increased in order to provide more reliable assessment of the clustering algorithm efficiency.

### 8. REFERENCES

[1] O. Andersen, P. Dalsgaard and W. Barry, 
Data-Driven Identification of Poly- and Mono-phonemes for four European Languages. 1993, Proc. EUROSPEECH '93, Berlin, pp. 759 - 762


