Unsupervised Adaptation for Acoustic Language Identification

Ekaterina Timoshenko, Josef G. Bauer

Siemens AG, Corporate Technology, Otto-Hahn-Ring 6, 81730 Munich, Germany
{Ekaterina.Timoshenko.ext,Josef.Bauer}@siemens.com

Abstract

Our system for automatic language identification (LID) of spoken utterances is performed with language dependent parallel phoneme recognition (PPR) using Hidden Markov Model (HMM) phoneme recognizers and optional phoneme language models (LMs). Such a LID system for continuous speech requires many hours of orthographically transcribed data for training of language dependent HMMs and LMs as well as phonetic lexica for every considered language (supervised training). To avoid the time consuming process of obtaining the orthographically transcribed training material we propose an algorithm for automatic unsupervised adaptation that requires only raw audio data as training material covering the requested language and acoustic environment. The LID system was trained and evaluated using fixed and mobile network databases (DBs) from the SpeechDat II corpus. The baseline system – based on supervised training using fixed network databases and covering 4 languages - achieved a LID error rate of 6.7 % for fixed data and 19.5 % for mobile data. Using unsupervised adaptation of the HMMs trained on fixed network data the error rate for mobile DBs database mismatch is reduced to 10.6 %. Exploring a situation when orthographically transcribed training data is not available at all multilingual HMMs were unsupervised adapted to fixed and mobile DBs and perform at 10.8 % and 12.4 % error rate respectively.

Index Terms: Language Identification, unsupervised adaptation, Parallel Phoneme Recognition, Hidden Markov Model.

1. Introduction

Recent LID systems ([1], [2]) are designed to identify a language of a spoken utterance from a given set of languages to be recognized. Different LID algorithms have different requirements for the training data which is processed during training to produce corresponding models. For a real use case the problem is often either the availability of suitable models or the availability of orthographically transcribed training material that can be used to create them. Obtaining such a training material is time consuming and expensive and it limits the application scope of PPR approach.

Using the PPR-based LID system requires for every language to be recognized the existence of available orthographically transcribed training material and phonetic lexicon that should be fit together. In this paper we show how the high development cost of a PPR-based LID system can be overcome by using unsupervised adaptation techniques. Unsupervised training of acoustic models has been already successfully applied for speech recognition tasks [3]. Using unsupervised adaptation of acoustic models for a language identification task can give even more advantages.

The underlying idea is to use a phoneme recognizer based on already existing acoustic model to transcribe data that later will be applied for an adaptation. The unsupervised adaptation of acoustic models can be used in different real world scenarios: for adaptation of already existing language specific models to particular database or for adaptation of multilingual models for a new language that was not used for training. The corresponding LMs can be estimated by computing statistics over the phoneme sequences produced by the adapted phoneme recognizers.

The proposed LID system based on parallel phoneme recognition is described in the next section. Section 3 presents unsupervised learning algorithms for acoustic model adaptation and estimating LMs. Section 4 gives a description of SpeechDat II corpus used for training the LID system and its evaluation. Then we provide the results for the LID system based on language specific models and models obtained using unsupervised adaptation algorithms. Experiments with different sizes of training material investigate the sufficient amount of adaptation data.

2. Description of the LID system

The LID system is designed using the PPR approach that requires a language-dependent phoneme recognizer for every language in the given set. All phoneme recognizers use CDHMM theory for acoustic modeling and were implemented using a toolkit for HMM-based automatic speech recognition created for Siemens ASR Technology [4]. Every CDHMM uses 2048 Gaussians and monophone models. An artificial neural network is used as an additional component of the LID system. The ANN is implemented as three layer perceptron with ten hidden nodes and the number of input and output nodes being the number of considered languages.

The architecture of the proposed LID system is presented in Fig. 1. An acoustic preprocessing component of the system re-

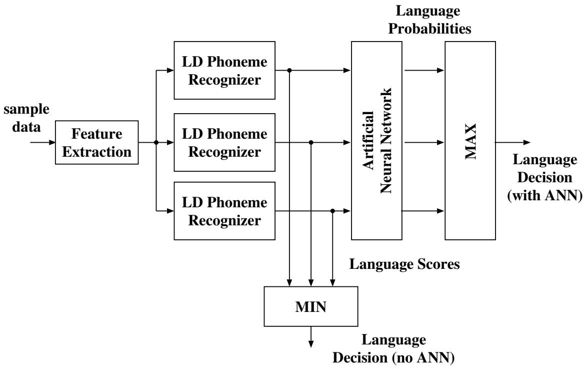

Figure 1: *LID system design*

ceives an input acoustic waveform and converts it into a set of feature vectors. Then the classification is organized as follows:

- The set of feature vectors proceeds through all phoneme recognizers in parallel.
- LMs are used as integral parts of the phoneme recognizers; they are estimated with maximum likelihood objective.
- Each phoneme recognizer finds the most likely phoneme sequence based on first best Viterbi decoding and calculates a neglog likelihood score for it.
- The system decision is made either by taking a minimum of the normalized language scores (here the sum of the minimal neg-log state specific likelihoods is subtracted and the result is divided by the number of frames), or by using an artificial neural network with language scores as an input in order to extract a-posteriori probabilities for every language in the set.

The corresponding ANN is trained on normalized language-dependent neglog-likelihood scores produced by processing the training material through the phone recognizers. During the training the ANN takes the scores as an input and binary values as an output. The output node that corresponds to the spoken language gets ”1”, otherwise it gets ”0”. By iterating the learning procedure, the ANN weighting coefficients are estimated.

During the classification the ANN gets normalized scores as an input and aims to produce a-posteriori probabilities for every language. Then the system hypothesizes a language with maximum a-posteriori probability. Using such kind of ANN does not require orthographically transcribed language specific databases for the ANN training and thus does not increase system training costs.

In contrast with other LID approaches the PPR-based system allows the phoneme recognizer to use the language-specific phonotactic constraints during the Viterbi decoding process so that the joint acoustic-phonotactic likelihood of the phoneme sequence is computed. Thus, the most likely phoneme sequence is optimal with respect to the combination of both acoustics and phonotactic information.

3. Unsupervised adaptation

For many audio data sources there are no corresponding orthographical transcriptions and phonetic lexicons which are necessary for supervised training of a LID system. In such situations it is possible to use unsupervised learning — meaning that the system does not receive the orthographical transcriptions of the training data and does not require the existence of a phonetic lexicon. Instead it establishes the signal-phoneme correspondence itself based on the statistical regularities of the data. Unsupervised training algorithms can be used for acoustic models adaptation and for obtaining LMs.

3.1. Unsupervised adaptation of acoustic models

The adaptation of acoustic models using non-orthographically transcribed training material in an unsupervised manner can be applied in the following real world situations:

- Database / environment adaptation: language specific models are avaliable for all languages in the set but they are trained on another database / environment.

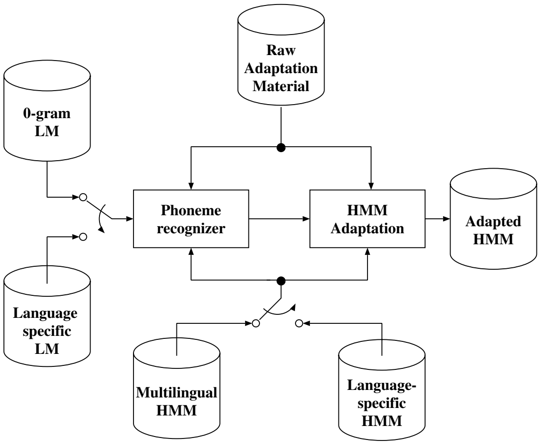

Figure 2: *Unsupervised HMM adaptation*

- Language adaptation: language specific models are not avaliable.

A block diagram of the unsupervised adaptation process is presented in Fig. 2. Non-orthographically transcribed training material passes through a phoneme recognizer based on existing initial model. The initial acoustic model (language specific or multilingual HMM) in combination with LM (language specific bigram or 0-gram) is used to transcribe the training material. Signal-phoneme correspondences produced by a phoneme recognizer are used to adjust the initial model to a specific database or a particular language depending on the particular task.

The adaptation method uses a combination of Maximum Likelihood Linear Regression (MLLR) with one global fully assigned MLLR matrix and Maximum A-Posteriori (MAP) adaptation. This approach was found to be effective, stable and flexible for open vocabulary adaptation [5]. Combining both methods the mean vectors are reestimated as follows. MAP is applied for every mean vector that was seen during the adaptation process. For all other mean vector the MLLR matrix is applied.

The performance of the models adapted in such a way also depends on the characteristics of the adaptation material. For adaptation it is better to use phonetically rich sentences that cover all phonemes used in the HMM. Typically the words in the training material are different from the words in the language identification task.

3.2. Unsupervised learning for language models

An unsupervised method can be also used to create new LMs that are integrated to the phoneme recognizers as it is usually done for back-end in LID systems that are based on the Phone Recognition followed by Language Modeling (PRLM) approach [2].

In this case the language specific acoustic models cannot be trained and the existence of a multilingual HMM is required. This HMM first should be adapted to the particular language as described above. Then the raining material is passed through a corresponding phoneme recognizer using existing adapted multilingual acoustic model. The phoneme sequences produced by the recognizer are used to compute the speech statistics necessary for LM estimation.

	Training	Development	Test
Mean utt. length (sec)	7	7	7
Data amount (hours)	124	14	5

Table 1: *Amounts of speech data from fixed network databases*

	Training	Test
Mean utt. length (sec)	8	8
Data amount (hours)	70	5.5

Table 2: *Amounts of speech data from mobile network databases*

4. Speech corpus

For the training and evaluation of the LID system described in this work we used the SpeechDat II ([6], [7]) corpus. In particular, for all experiments we used databases from fixed and mobile telephone networks for the following languages: German (DE), English (EN), Italian (IT) and Dutch (NL).

Both groups of databases were divided into several subsets using speakers defined in SpeechDat II for training and testing so that they do not overlap and do not contain utterances with common wordings as it is done in [7]. The training sets were used to train and adapt the HMMs and to estimate the LMs. A development set of fixed DBs was used for training the ANNs for language classification. Finally, the evaluation was performed on the test sets. The information about fixed and mobile network databases used in this work is contained in tables 1 and 2 respectively.

All sets contain only phonetically reach sentences. The numbers of utterances for all languages in each described set were approximately balanced.

5. Experiments and results

To examine the performance of the LID system language specific HMMs for four languages (see previous section) were trained using the orthographically transcribed training material from the fixed network databases. The LMs were created for every language by computing the statistics for every bigram in the phoneme sequences provided by the orthographically transcribed utterances from the training set of fixed DBs. The weights for corresponding ANN were optimized on the development set of fixed network DBs.

Trained in such a way the LID system was evaluated on both fixed and mobile test sets. The average error rates for the fixed and mobile databases come up to 6.7 % and 19.5% respectively (see tables 3 and 4 for language specific error rates). The performance of the system on mobile DBs is much worse than the results for fixed DBs, probably due to the fact that the speech material for training and testing the system is taken from different databases.

5.1. Database adaptation

To confirm an influence of differences between the databases that are used to train the system and those on which the system is tested (database mismatch) another neural network was created. The ANN with the same topology as the previous one was trained on the subset of mobile training material. For absolute agreement of system configurations the training material for ANN was composed of lists that contain the same number of utterances per language as it was used for fixed DBs. The mean length of the utterances of fixed and mobile DBs were nearly the same.

As expected, the system error rate of 15.4 % is better than with previous ANN. This shows that the system is quite sensitive to the database mismatch. To overcome this disadvantage the language specific HMMs trained on fixed DBs were adapted to mobile DBs in the unsupervised manner. The average error rate for resulting LID system was reduced to 10.6 %. Consequently, unsupervised database adaptation can significantly improve system performance.

For the unsupervised adaptation in the experiment above the whole available amount of training material was used. To explore the influence of the size of the training set used for adaptation on the resulting system performance a sequence of analogous experiments with varying sizes of mobile training set was performed. The size of the training set means the total number of utterances coming from DE, EN, IT and NL training sets where all languages are presented by nearly equal amounts of speech.

Starting from 31000 of training utterances the size of the training set was stepwise reduced to 100 utterances. Every set, consisting of the utterances randomly taken from the initial one, was used to adapt HMMs to mobile DBs. The adapted HMMs were used to train the corresponding ANNs. The resulting systems were tested, their performances are presented in Fig. 3. The system performance up to a training set of 1000 utterances does not change significantly (size of 1000 means 250 utterances per language which corresponds to half an hour of speech). Even the adaptation set of size of 100 utterances can be useful. So, 1000 utterances or 2.22 hours of speech is a sufficient amount of training material for the LID system adaptation.

5.2. Adaptation of multilingual models

In order to test the unsupervised adaptation of multilingual HMM to a particular language we assume that the language specific knowledge sources are not available for DE, EN, IT and NL. Assume also, that training data is available for French (FR), Polish (PL) and Spanish (ES) languages of fixed network databases. So, the multilingual HMMs were trained on FR, PL and ES ortho-

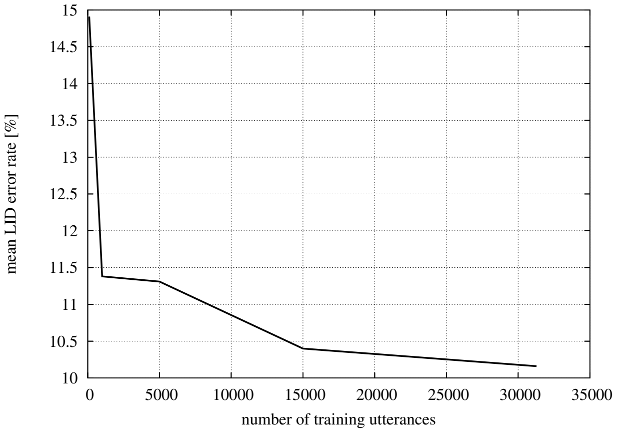

Figure 3: *Dependence of the system performance on the size of training set (10000 utterances equivalent to 22.2 hours of speech)*

Models	ER in %				
	DE	EN	IT	NL	Mean
sup. learning (HMMs, LMs), ANN	7.6	7.8	5.4	5.9	6.7
unsup. learning (HMMs, LMs), ANN	13.7	15.2	7.0	7.4	10.8

Table 3: *Performance of different LID systems for fixed DBs*

Models	ER in %				
	DE	EN	IT	NL	Mean
sup. learning (HMMs, LMs), ”fixed” ANN	20.3	18.1	4.9	42.7	19.5
sup. learning (HMMs, LMs), ”mobile” ANN	12.7	22.4	12.3	14.61	15.4
unsup. adaptation of lang.spec. HMMs, sup. learning of LMs, ”mobile” ANN	11.3	14.4	6.4	11.6	10.9
unsup. learning (mult. HMMs, LMs), ”mobile” ANN	14.7	10.3	9.3	17.1	12.4

Table 4: *Performance of different LID systems for mobile DBs*

graphically transcribed training sets with resulting 46 phones. Utilizing the unsupervised adaptation scheme the multilingual HMMs were adapted to DE, EN, IT and NL languages using training material from fixed DBs orthographically transcribed with the help of multilingual phoneme recognizer. The corresponding LMs was created also using the unsupervised learning technique: first, the training data was transcribed with the help of the phoneme recognizers based on the adapted multilingual HMMs, and then the produced phoneme sequences were used to obtain LMs. The resulting LID system was evaluated on fixed and mobile databases.

Table 3 gives the comparison of the error rates (ER) for the LID systems based on the models estimated by supervised and unsupervised training. The absolute difference in mean error rates comes up to 4 %. However, in case of lack of orthographically transcribed training material the unsupervised training can be used to create an appropriate LID system.

The results over all experiments performed for mobile databases and described above are presented in Table 4 that shows how the error rates of the LID system may vary depending on the availability of training material. The best results are obtained by the system based on unsupervised adaptation of language specific models. Concerning the Table 4 we conclude that in case of a database mismatch the performance of the system that utilize the unsupervised learning of multilingual HMM using training data without orthographical transcription is relatively 36 % better than performance of language specific system.

6. Conclusions

In this work we have shown that the most expensive requirement of PPR-based LID systems (existence of orthographically transcribed training material for every language in the task) can be overcome by using the unsupervised adaptation techniques. Experiments have demonstrated that phoneme recognizers can be effectively used to transcribe data for unsupervised learning and adaptation of the LID systems. According to the different real world scenarios we propose the following applications of unsupervised learning for the LID task:

- adaptation of an existing language specific HMM to a particular database in order to overcome the mismatch between training and test databases;
- adaptation of a language independent (multilingual) HMM to a language of interest that can easily extend already existing system to new languages.

The importance of the unsupervised learning for the LID task is proven by the results of HMM adaptation together with LMs estimated using unsupervised learning which are comparable with the performance of the LID system based on the language specific HMMs and LMs. An amount of 250 utterances (or half an hour of speech) per language used as training material was found to be sufficient for the model adaptation of the LID system.

7. References

[1] Muthusamy, Y.K., Berkling, K., Arai, T., Cole, R.A. and Barnard, E., ”A Comparison of Approaches to Automatic Language Identification Using Telephone Speech”, EUROSPEECH 1993, 1:1307-1310.

[2] Zissman, M.A., ”Comparison of Four Approaches to Automatic Language Identification of Telephone Speech”, IEEE Trans. Speech and Audio Proc., 4(1):31–44, 1996.

[3] Lamel, L., Gauvain, J.-L. and Adda, G., ”Unsupervised Acoustic Model Training”, ICASSP 2002, I:877–880.

[4] Varga, I., Aalburg, S., Andrassy, B., Astrov, S., Bauer, J.G., Beaugeant, C., Geissler, C. and Höge, H., ”ASR in Mobile Phones — an Industrial Approach”, IEEE Trans. Speech and Audio Proc., 10(8):562–569, 2002.

[5] Bauer, J.G., ”A Study on Open Vocabulary Speaker Adaptation for Low Footprint Speech Recognition”, Proc. Advances in Speech Technology, International Workshop, Maribor, Slovenia, 2004.

[6] ”ELRA web site”, http://www.elra.info.

[7] Caseiro, D. and Trancoso I.M., ”Spoken Language Identification Using The SpeechDat Corpus”, Proc. Int. Conf. on Spoken Language Processing (ICSLP), 1998.