Using Speech Rhythm for Acoustic Language Identification

Ekaterina Timoshenko, Harald Höge

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

Abstract

This paper presents results on using rhythm for automatic language identification (LID). The idea is to explore the duration of pseudo-syllables as language discriminative feature. The resulting Rhythm system is based on Bigram duration models of neighbouring pseudo-syllables. The Rhythm system is fused with a Spectral system realized by parallel Phoneme Recognition (PPR) approach using MFCC's. The LID systems were evaluated on a 7 languages identification task using the SpeechDat II databases. Tests were performed with 7 seconds utterances. Whereas the Spectral system acting as a baseline system achieved an error rate of 7.9 % the fused system reduced the error rate by 10 % relatively.

Index Terms: language identification, speech rhythm, pseudo-syllables, duration

1. Introduction

The aim of an Automatic Acoustic LID system is to determine the language from a given utterance. Progress in LID technology has been pushed considerably by the evaluation campaigns performed by NIST [1]. The work of this paper is performed in the spirit of the NIST evaluation paradigm but it uses an own evaluation dataset based on the languages of the widely used SpeechDat database family [2] designed to train commercial speech recognizers. Due to this design these databases are very well suited to investigate LID systems [3]. This family covers a wide range of languages which is continuously extending (see for example LILA project [2]) and allows to evaluate LID approaches for various language sets and to study language specific properties. The set of possible languages to be identified can be closed or open. Due to the focus of the paper we restrict our investigations on the closed set problem.

Standard LID systems [4] are based on the acoustic-phonotactic information and consider a phonetic modeling system producing a sequence of phones that are evaluated using language specific statistical models. Such LID systems use spectral features as the MFCC's for acoustic modeling.

Recently prosodic features as pitch, intonation, rhythm have been regarded to improve further LID systems [5, 6, 7]. It is well known, that the basic prosodic features, i.e. pitch and duration, are hard to extract reliably especially in noisy environment [8]. To use prosodic features for LID basically three problems have to be solved:

- find a suited model for using the prosodic features,
- extract reliable the basic prosodic features,
- find a suited method to fuse prosodic and spectral information.

Starting with a MFCC based PPR system described in [9] we investigate the use of one of the prosodic feature - speech rhythm.

In the following sections the paper is organized as follows. Section 2 concerns the modeling of the prosodic feature - rhythm and our approach to extract the basic prosodic parameter - duration. Section 3 describes our new LID system fusing spectral and rhythm information. The last two sections are dedicated to the evaluation dataset and the presentation of experimental results.

2. Rhythm and Duration

2.1. Modeling of Rhythm

Although speech rhythm was studied by researchers for decades using rhythm for LID is not straightforward. Traditionally languages are classified into three rhythmic classes namely stress-timed, syllable-timed and mora-timed based on the isochrony theory which is defined as the property of speech to organize itself in pieces equal or equivalent in duration. This rhythm class hypothesis contradicts with more recent experiments which brief discussion can be found in [10, 11] and assumes that rhythmic differences between languages can be explained by the syllable structure and the presence (or absence) of vowel reduction. Nevertheless it is common sense that rhythm is related to the duration of some speech units.

At the same time there are still several open questions concerning an appropriate treatment of speech rhythm for language identification, namely

- how to segment speech into suited rhythmic units, and
- develop a language independent approach for rhythm modeling.

Here we concentrate on the investigation of syllables as more intuitive rhythmical units and claim that speech rhythm can be modeled by the durations of two successive syllables. In [12] it was already shown that rhythmic events can be described by two-dimensional scatter plots spanned by the duration of successive syllables. This method is able to show the clear distinctions between stress-timed and syllable-timed languages. These results were obtained by the analysis of duration values given by manual segmentation of speech into syllables.

For LID an automated segmentation into phonetic units is needed. Syllables are not very suited for language independent approach, because the set of syllables has to be provided for each new language investigated. But most important the boundaries of syllables are not easy to detect with the needed precision, because they are often found in between of consonant clusters, where automatic segmentation is difficult. Instead of syllable we use the notion of pseudo-syllable first introduced in [13]. According to [13], the most frequent syllable structure over world's languages is CV structure, where C is a consonant and V is a vowel. The pseudo-syllable is defined as a pattern C^nV with n an integer that can be zero. So in this case the automatic language independent algorithm for rhythm feature

extraction can be easily derived using a consonant-vowel segmentation mechanism. This approach needs no language specific knowledge concerning syllables and segment boundaries are always at vowels, which can be detected more reliable.

Unlike [13], where the C^nV structure of pseudo-syllable was used as a rhythm feature, here the idea is to explore the potential of the duration of pseudo-syllable itself as feature for LID. We characterize a pseudo-syllable by the total duration consisting of its consonant and vowel parts.

A speech utterance is modelled as a sequence of pseudo-syllables $S = s_1 s_2 \dots s_N$. Assigning to each pseudo-syllable s_j its duration d_{s_j} for each language L_i $(i = 1, 2, \dots, M)$ we model rhythm via a Bigram duration model $P(d_{s_j}, d_{s_{j+1}} \mid L_i)$.

For a given sequence of pseudo-syllable durations $D = d_{s_1}, d_{s_2}, \dots, d_{s_N}$ the rhythm language score $s_R(D \mid L_i)$ is given by

$$s_R(D \mid L_i) = -\sum_{j=1}^{N-1} \log P(d_{s_j}, d_{s_{j+1}} \mid L_i) \qquad (1)$$

In our LID system segmentation is performed with a language independent analysis of the signal, which provides for each utterance a single estimated sequence D. The resulting score $s_R(D \mid L_i)$ need not to be normalized when scores derived from the same number of estimated pseudo-syllables are compared.

Instead of the joint propabilities $P(d_{s_j}, d_{s_{j+1}} \mid L_i)$, the conditional distributions that lead to an approximation

$$P(D \mid L_i) \approx P(d_1) \prod_{j=1}^{N-1} P(d_{s_j} \mid d_{s_{j+1}}, L_i) \qquad (2)$$

could be used. Although this approach would model the whole timing structure of the sequence of pseudo-syllables, experimentally we found that Rhythm system based on the joint probabilities gives better results.

The language is classified either with a pseudo maximum likelihood method:

$$\hat{L} = \underset{L_i}{\operatorname{argmin}}\, s_R(D \mid L_i), \qquad (3)$$

where $s_R(D \mid L_i)$ is defined by Equation (1) or by an Artificial Neural Network (ANN) as described in Section 3.

2.2. Extraction of pseudo-syllable durations

To determine the rhythm score the durations of the pseudo-syllables are needed. The basic idea is to detect the temporal ending point of each vowel and measure the time elapsed between successive vowel ending points. For vowel detection we considered two methods:

- Signal based methods,
- Hidden Markov Model (HMM) based methods.

In the signal based methods spectral energies are extracted from speech, where energy peaks are used to detect the position of the vowel [14]. With this method the ending point of a vowel is hard to detect.

The second approach is based on using an HMM based phoneme recognizer which outputs not only the sequence of phonemes recognized but also the timing of the state sequence given during the Viterbi decoding process. The temporal ending points of the vowels are determined by the timing of the last state of each vowel.

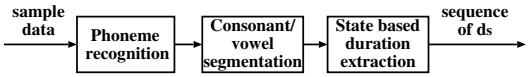

Figure 1: *Extraction of pseudo-syllable durations*

Initial experiments using language specific HMMs showed, that the HMM based method leads to more accurate detection of the vowels. So we decided to use the HMM based method to determine the duration. As we want to implement a language independent approach we used a multilingual HMM where phonemes are modelled with 6 states, 2 of which are tied. The accuracy of obtained durations is limited by the time shift value that in our case is equal to 15 milliseconds.

The rhythm feature extraction procedure is performed in several steps as it is presented in Fig. 1:

- The input signal is transformed into a sequence of phonemes with corresponding phoneme durations using a language independent phoneme recognizer realized by a multilingual HMM.
- The phoneme sequence is converted into a consonant-vowel sequence.
- According to the notion of the pseudo-syllable the consonant-vowel sequence is parsed into patterns matching the structure C^nV. For the resulting sequence of pseudo-syllables the corresponding durations d_s are computed.

The sequences of pseudo-syllable durations extracted as described above are used to estimate the language specific rhythm models by calculating the probabilities for all possible pairs of pseudo-syllable durations. The probability distributions of duration are given for discrete duration values, which are determined by the numbers of frames regarded. The discrete distribution values building a histogram are not smoothed. For unseen durations a fixed floor value is used.

To illustrate the rhythm models we plot the corresponding distributions on the three dimensional space where the x-axis presents the duration of a pseudo-syllable i in frames, the y-axis presents the duration of subsequent pseudo-syllable $i+1$ and z-axis presents the probability for that pair. As an example of such

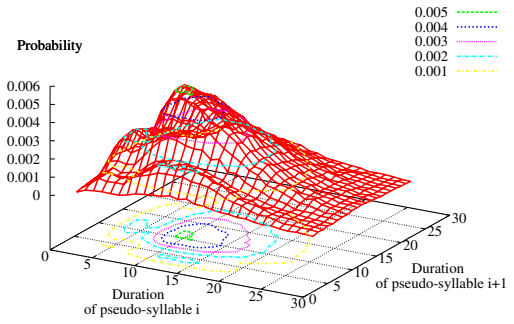

Figure 2: *Probability distribution for German language obtained by multilingual phoneme recognizer*

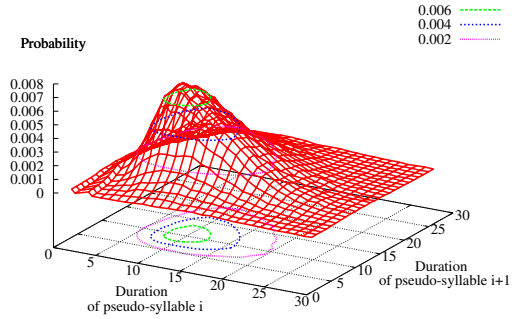

Figure 3: *Probability distribution for German language obtained by forced Viterbi algorithm*

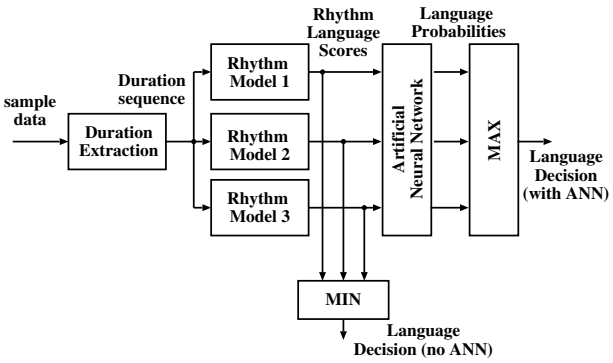

Figure 4: *Rhythm system design for 3 languages identification task*

plots the Fig. 2 presents the distribution for German language evaluated with a multilingual phoneme recognizer. Here the curves on the x-y surface show the contours for the different probability values.

To estimate the accuracy of the pseudo-syllable durations obtained in such way we performed an experiment using forced Viterbi algorithm for the same data with orthographical transcriptions that gave us the actual durations. The results are plotted in Fig. 3. The differences between two plots presented in Fig. 2 and 3 can be explained not only by the accuracy of segmenting speech during acoustic processing but also by the relatively low phoneme recognition rate of the multilingual HMM used in this work which is in the range of 22 % only.

3. Description of the LID system

3.1. Spectral System

The LID system called here the Spectral system [9] is based on the evaluation of likelihood scores provided by language specific phoneme recognizers running in parallel (PPR approach). The recognizers use MFCC's as spectral features and intergated bigram phoneme models. For every frame f_k each phoneme recognizer delivers a likelihood score $-\log P(x_k \mid Q_l, L_i)$, where Q_l denotes an HMM state of the optimal path found by Viterbi search. For the sequence of frames $F = f_1, \ldots, f_K$ the spectral language score is defined as

$$s_S(F \mid L_i) = -\sum_{k=1}^{K} \log P(x_k \mid Q_L, L_i)) \qquad (4)$$

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. The ANN is implemented as three layer perceptron with ten hidden nodes, the number of input and output nodes being the number of considered languages. The ANN is trained on the z-normalized scores produced by processing the training material through the recognizers and uses sigmoid function for activation and Backpropagation as a learning algorithm. During the classification the ANN gets normalized language scores as input and aims to produce a-posteriori probabilities for every language. The system hypothesizes a language with maximum a-posteriori probability.

3.2. Rhythm System

During the recognition phase, for a sequence of pseudo-syllable durations computed as described in Section 2.2 and presenting a test input utterance, the rhythm scores for all languages are calculated and the language with minimal score is hypothesized. The Rhythm LID system is presented in Fig.4. Since in [9] we have found the great advantage of a neural network on the recognition results, the Rhythm system has an postprocessing ANN as an additional optional classifier. The ANN is trained on the rhythm scores and is working as described in the previous section.

3.3. Fused LID System

The resulting language scores coming from the Spectral and the Rhythm systems are combined using an ANN. The ANN is implemented as before with only difference in the number of input nodes which is equal to the double amount of languages to identify. The ANN gets normalized language scores form the Spectral and Rhythms systems as input and extract the corresponding a-posteriori probabilities. The fused LID system is presented in Fig. 5.

4. Speech Corpus

For the training and evaluation of the LID system we used the SpeechDat II ([2]) corpus consisting from fixed telephone network recordings. The target languages are German, French, English, Spanish, Italian, Dutch and Polish.

The data for every language was divided into several subsets using speakers defined in SpeechDat II for training, devel-

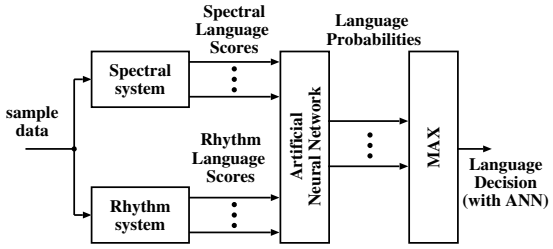

Figure 5: *Fusion of LID systems*

Systems	ER in %							
	German	English	Spanish	French	Italian	Dutch	Polish	Mean
Spectral	8.47	15.67	10.89	4.5	16.76	5.61	8.0	9.98
Spectral+ANN	7.64	7.67	7.68	4.75	12.25	6.21	8.53	7.92
Rhythm	79.73	74.17	55.54	60.0	80.36	90.45	82.53	74.68
Rhythm+ANN	61.63	67.83	45.71	86.5	80.9	82.58	60.0	68.43
Spectral+Rhythm	8.14	8.17	6.79	3.75	10.45	5.3	6.53	7.1

Table 1: *Performance of different LID systems*

opment and testing so that they do not overlap and do not contain utterances with common wordings as it is done in [3]. The amounts of data per language everage 30 hours of speech for HMM and rhythm models training and 3.5 hours for ANN training. For every language about 600 test utterances were used. Every utterance is about 6-7.5 seconds which corresponds to 12-20 pseudo-syllables. All sets contain only phonetically rich sentences.

5. Experiments and Results

To examine the performance of proposed approach for the Spectral system language specific HMMs and LMs for every language (see previous section) were trained using the orthographically transcribed training material. The average LID error rate (ER) for the Spectral system comes up to 9.98 % without neural network and 7.92 % with ANN.

For the Rhythm system the multilingual HMM used for pseudo-syllable duration extraction was trained on all avaliable training data. The rhythm models were created by computing the histogram statistics for every pair of pseudo-syllables provided by training data.

Trained in such a way Rhythm system was first evaluated independently. For the 7 languages task the system performs the 74.68 % of error rate. Using the ANN the error rate was reduced to 68.43 %. In spite of the high error rate of the Rhythm system itself, the fusion of both systems outperforms the best results of the Spectral system by 10 % relatively. Table 1 displays the language specific as well as mean error rates for the different LID systems.

6. Conclusions and Future Work

In this paper we provided a first step toward using rhythm as an additional source of information for LID task. The experimental results show that the speech rhythm presented by sequence of pseudo-syllable durations can be successfully used for LID.

At the same time there are plenty possibilities to improve the achieved results. First of all, there is a necessity for the increasing the accuracy of the consonant-vowel segmentation algorithm coming from the differences between probability distributions presented in Fig. 2 and 3. This can be done either by improving the quality of the multilingual phoneme recognition, or by implementing a new language-independent vowel/non-vowel detection scheme.

Another way for improvement is contained in the modeling the Bigrams for pseudo-syllable duration. The probabilities of the rhythm parameters can be represented by Gaussian Mixture Models or simply approximated by a continuous function.

In the future we are also going to try to find more appropriate fusion algorithm or to implement the normalization procedure with respect to speaking rate.

7. Acknowledgements

We would like to thank Josef G. Bauer for the training of language specific and multilingual HMMs used in this work.

8. References

[1] "NIST Language Recognition Evaluation", http://www.nist.gov/speech/tests/lang/index.htm.

[2] "SpeechDat web site", http://www.speechdat.org.

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

[4] 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.

[5] Tong, R., Ma, B., Zhu, D., Li, H. and Chng, E.-S., "Integrating acoustic, prosodic and phonotactic features for spoken language identification", Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2006.

[6] Lin, C.-Y. and Wang, H.-C., "Fusion of phonotactic and prosodic knowledge for language identification", Proc. Int. Conf. on Spoken Language Processing (ICSLP), 2006.

[7] Rouas, J.-L., "Modeling Long and Short-term prosody for language identification", Proc. Int. Conf. on Spoken Language Processing (ICSLP), 2005.

[8] Höge, H., "Basic Parameters in Speech Processing. The need for evaluation", To be published in Archives of Acoustics, Vol. 32, 2007.

[9] Timoshenko, E. and Bauer, J.G., "Unsupervised Adaptation for Acoustic Language Identification", Proc. Int. Conf. on Spoken Language Processing (ICSLP), 2006.

[10] Grabe, E. and Low, E.L., "Duration variability in speech and the rhythm class hypothesis", Papers in Laboratory Phonology 7, 2002.

[11] Ramus, F., Dupoux, E. and Mehler, J., "The psychological reality of rhythm classes: perceptual studies", The 15th International Congress of Phonetic Sciences, 2003.

[12] Wagner, P., "Visualization of Speech Rhythm", To be published in Archives of Acoustics, Vol. 32, 2007.

[13] Farinas, J. and Pellegrino, F., "Automatic rhythm modeling for language identification", Proc. Eurospeech Scandinavia, 2001.

[14] Howitt, A.-W., "Automatic syllable detection for Vowel Landmarks", PhD thesis, Massachusetts Institute of Technology, 2000.