AN EFFICIENT PHONOTACTIC-ACOUSTIC SYSTEM FOR LANGUAGE IDENTIFICATION

Jiří Navrátil and Werner Zühlke

Department of Communication and Measurement
Technical University of Ilmenau, P.O.Box 100565, 98684 Ilmenau, Germany
e-mail: jiri.navratil@e-technik.tu-ilmenau.de

ABSTRACT

This paper presents a combined two-component system for language identification based on phonotactic and acoustic features. The phonotactic part consisting of a multilingual phone-recognizer with a double bigram-decoding architecture and a phonetic-context mapping is supported by a second part with pronunciation modeling of the recognized phone-sequence using Gaussian density models. Both parts are post-processed by a neural-based final classifier. Measured on the NIST'95 evaluation set, the described system outperforms state-of-the-art components and, at the same time, requires considerably less computational expense, as compared to implicit phonotactic-acoustic modeling and parallel recognizer architectures.

1. INTRODUCTION

With the trend in globalizing the communication technology and providing services to a wide, multilingual public, the ability of machines to distinguish between languages has become increasingly important. Automatic language identification (ALI) finds its application potential in multilanguage spoken dialog systems, such as information terminals, databases, or archives, as well as in the human-human communication (call-routing, automatic translation).

During the past decade intensive research efforts have been made covering different solutions for ALI. Several sources of language-discriminative information have intuitively been addressed as relevant for this task: the prosody, the acoustics, and the grammatical and lexical structure.

Besides prosodic and acoustic features [1], a very promising and feasible way of acquiring language-specific information is the modeling of statistical constraints inherent in phonetic chains - the phonotactics. In this sense phonotactics can be viewed as a subset of grammatical and lexical rules of a language. Several contributions were published dealing with the use of phone n-grams, particularly bigrams, for modeling and classifying languages [2], [3]. Various configurations of multiple language-dependent phone-recognizers, run in parallel, were designed to better represent the phone repertoire and to improve the performance of simple phonotactic components. In [4] a double-bigram decoding architecture was introduced employing a single multi-language decoder with separate sets of language models within and outside the phone-recognizer, which outperformed the parallel architecture and reduced the computational expense.

Despite the high efficiency of the phonotactic features it seems obvious that only a way of incorporating multiple sources of knowledge will lead to the robustness necessary for practical applications. In [2] taking the phone durations proved to increase the identification accuracy. Another combination, namely implicit phonotactic and acoustic modeling, was done in [3] by using several language-dependent phone-recognizers with implicit language models where the resulting acoustic likelihoods were taken for the final classification. In both cases the computational costs were considerable due to the multiple recognition process.

This contribution presents recent development of our ALI system [5] toward a multi-approach solution that exploits acoustic pronunciation differences between languages by Gaussian probability density models and combines these features with the phonotactics using a neural-based classifier. Hereby, the system is considered as effective in terms of improved performance and less computational expense, as compared to parallel decoder architectures.

In section 2 a description of the phonotactic component serving as the baseline system in this work is given. The acoustic features and models are dealt with in section 3. Subsequent sections detail on the final classifier, the databases, and give the experimental results obtained with the new system.

2. PHONOTACTIC COMPONENT

As the phonotactic component the system described in [5] was employed. Here, a multilingual phone-recognizer and a double bigram-decoding architecture applied, as shown in Figure 1 (Block 1). During the Viterbi decoding process, M (=6) language-dependent bigrams are used to weight the transitions between individual phones thus generating M phone-streams. With each stream an independent set of N language models is connected. Resulting scores are combined together and fed to the final classifier. The bigrams used within the Viterbi-decoder were estimated on original transcriptions in six languages, whereas the language models were trained on the corresponding decoded phone-streams. This proved to outperform the parallel-decoder systems and was, at the same time less computationally expensive as the decoder can carry out a synchronous Viterbi pass for all streams [4].

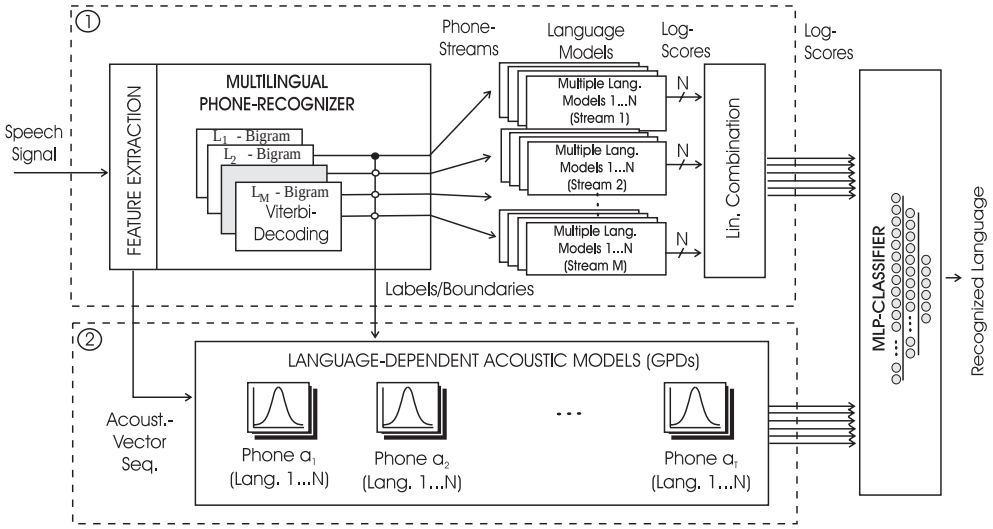

Figure 1: ALI system overview: (1) - Phonotactic block, (2) - Acoustic block.

Further on, for language modeling standard bigrams were combined with additional models to capture a wider phonetic context [5]. While the bigram acquires statistical dependencies of phone pairs, the so-called selection-matrix (SM) bigram gets information from phone triples, and a binary-decision tree (BT) model exploits constraints from up to three preceding phones. The additional models were shown to consistently improve the performance of the bigram models without the need for additional training data. The training procedures for both the SM and the BT models can be found in [5].

Given a spoken utterance decoded into M phone-streams $\overline{a}^{(l)} = a_1^{(l)}, ..., a_T^{(l)}$ $(l = 1, ..., M)$ the final phonotactic score of a language L_i $(i = 1, ..., N)$ is calculated as the sum over all stream-dependent scores:

$$S_{PT}(\overline{a} \mid L_i) = \\ = \sum_{l=1}^{M} S_{bi}(\overline{a}^{(l)} \mid L_i) + \alpha S_{sm}(\overline{a}^{(l)} \mid L_i) + \beta T(\overline{a}^{(l)} \mid L_i),$$

with S_{bi}, S_{sm}, and T denoting the log-scores for the bigram, the selection-matrix, and the binary-tree model respectively, and α, β being empirical weights. The resulting phonotactic scores $S_{PT}(\overline{a} \mid L_i)$ are fed to the final classifier.

3. ACOUSTIC COMPONENT

Due to the fact that the phone-recognizer is trained on multilingual data the phones are supposed to be recognized independently of the language being spoken. Thus, the acoustic differences of individual phones among languages can be acquired by modeling the language-specific pronunciation of each token from the phone-repertoire. A similar way to address the acoustic patterns was chosen in the segment-based approach by Hazen and Zue [6].

As depicted in Figure 1 the acoustic component (Block 2) consists of a set of language-dependent pronunciation models for each of the phones. During the classification, the decoded sequence together with phone boundaries is used for calculating acoustic language-hypotheses. For this, the original feature sequence is taken from the feature extraction block.

To calculate the score of a phone segment a certain number of feature vectors are cut out from the segment center (core vectors), averaged, and the resulting mean vector is put in the phone-dependent acoustic models of all languages. For a complete phone sequence the acoustic language score $\overline{a} = a_1, ..., a_T$ is calculated as follows:

$$S_{Ac}(\overline{a} \mid L_i) = \frac{1}{T} \sum_{t=1}^{T} \log \Pr(\overline{v}_t \mid a_t, L_i), \quad \text{for } i = 1, ..., N$$

whereby $\overline{v}_t$ denotes the acoustic evidence (mean vector) of the phone at the time t. The acoustic scores $S_{Ac}(\overline{a} \mid L_i)$ are then fed in the final classifier which combines them with the phonotactic component in a non-linear way. For modeling the acoustic patterns Gaussian probability density (GPD) models with an adaptive number of mixtures were employed.

The option to take a variable number of the core vectors from the phone segment (up to the complete segment) allowed to study the influence of coarticulations near the phone boundaries on the acoustic modeling (see Section 6).

It has to be noted that there are six different phone sequences emanating from the phone-recognizer which all supply possible phone boundaries for the acoustic component. The experiments, however, indicate that the performance of the acoustic models does not significantly depend on the stream used.

Optionally, also phone-dependent durations were included into the feature vectors thus extending the pure acoustic evidence by a first fundamental prosodic element.

4. FINAL CLASSIFIER

For the final classification including both the phonotactic and the acoustic scores a multi-layer perceptron with one hidden layer was applied to make the language decision. Perceptron classifiers are known to separate well non-linearly dependent information sources and proved to be superior to linear maximum-likelihood classifiers in our preliminary experiments.

5. IMPLEMENTATION

5.1. Databases

Up to nine languages from three multilingual speech corpora were used in the experimental work.

For training the phone recognizer, the phonotactic models, as well as the acoustic densities the OGI Multi-Language Telephone Speech Corpus [7] was taken, as described in [5]. Here, the acoustic models shared the training subset with the phonotactic language models (60 "45s-stories" per language). For training and cross-evaluating the neural-based final classifier the same languages from the new 22-language corpus collected at OGI [8] were used. The ca. 100 calls per language were processed by both system components whereby one half of the resulting scores served for training the network weights and the other half served for cross-validation. Calls from non-native speakers and those with a poor intelligibility or bad channel conditions were not included in the first half. Final results presented in this paper were obtained using the NIST[1] evaluation test set from March '95 which consisted of ca. 20 45-second phone calls and ca. 80 10s-excerpts from them in each language.

5.2. Training the acoustic models

For each phone a set of language-dependent GPD models with diagonal covariances was trained. K-means procedure was used to group the mean vectors of a phone into an optimal number of clusters from which the initial mixture parameters were estimated. Subsequently, the parameters were iteratively re-estimated according to the well-known Baum-Welch formulae. On average, there were seven mixtures per phone GPD. Complete vectors containing twelve Mel-warped cepstral coefficients, energy as well as their first derivatives were taken from the feature extraction and the cepstral mean substraction was carried out to compensate channel variations. Optionally, the segment duration was incorporated in the vectors as the 27-th feature dimension of the GPD.

6. EXPERIMENTS

Performance of the proposed system was tested using a closed set of *six* languages from the NIST evaluations involving English, German, Hindi, Japanese, Mandarin Chinese, Spanish, plus three other languages completing the *nine* language task: French, Tamil, and Vietnamese.

[1] National Institute of Standards and Technology

# of CV	1	3	7	14	ALL
Error rate	56.8%	56.9%	51.2%	51.2%	56.6%

Table 1: Acoustic performance versus number of the core vectors (CV) averaged

SD only	68.0%
7 CV	51.2%
7 CV+SD	49.8%
7 CV+SD+BW	48.9%

Table 2: Influence of the segment duration (SD) and Baum-Welch re-estimation (BW)

6.1. Acoustic experiments

In order to examine various parameter configurations, the acoustic component was evaluated in isolation first. All results were measured on the six-language-task with 10s-test utterances.

In the first experiment the influence of the phone-stream choice was analysed. The performance of the acoustic models was measured for each of the six phone-streams as well as with a special unigram-stream when no bigram was applied within the phonetic decoding. The results for all streams varied merely within a 4%-range. Thus the acoustic component can be viewed as nearly insensitive to the decoding stream chosen. For further examinations, only the stream 1 (English-bigram) was considered.

The acoustic patterns of the phones are known to vary with the phonetic context due to coarticulation effects. Ideal models should be context dependent and describe realisations with all possible contexts. Due to the sparseness of the data, however, the training must be restricted to context-independent models, i.e. phone realisations regardless of the context were taken for the parameter estimation. To examine the influence of the coarticulation variations of the vectors adjacent to the phone boundaries an experiment with variable number of feature vectors taken from the center of each phone was carried out. Table 1 shows the acoustic language error rate for different numbers of averaged core vectors (CV) ranging from 1 to complete segment (ALL). It can be seen that discarding the vectors near the phonetic boundary leads to better results due to reduced feature variations. On the other hand, taking too few vectors from the center results in a decreased robustness of the mean vector (1 and 3 CV). Seven core vectors were considered as sufficient and were taken for further experiments (for shorter phones all available vectors were used).

Results of another experiment, shown in Table 2, give the performance of the segment-duration feature (SD) alone and in combination with the complete acoustic vector which resulted in a slight improvement. After re-estimating the GPD models using the Baum-Welch formula (BW) the final error rate of the acoustic component decreased to 48.9%.

An interesting question was the relative importance of different phone classes in the overall acoustic perfomance.

Vowels	Others	Complete set
54.0%	51.7%	48.9%

Table 3: Relative performance of vowels and other phone classes

Configuration	6-Lang.		9-Lang.	
	10s	45s	10s	45s
AC-Component	48.9%	48.3%	51.3%	47.8%
PT-Component	12.8%	3.3%	22.6%	9.4%
Combined (MLP)	9.8%	0.8%	14.7%	5.6%

Table 4: Error rates on 10/45s utterances in the six- and nine-language-task (NIST'95)

In Table 3 error rates are given for cases where either (a) only vowels, or (b) all phones except vowels were used. Both groups in separation performed worse than using the complete phonetic repertoire.

Other accompanying experiments with the relative importance of *individual* phones indicated that the phone contribution to the overall accuracy varies from phone to phone and is dependent on the languages in the task. However, no improvement was achieved by excluding less relevant phones from the acoustic modeling.

6.2. Final Results

For evaluating the final system the acoustic and phonotactic scores were calculated for six and nine languages and normed within the range [-1,1] for the MLP-classifier. The training of the network was done on the two development sets as described in section 5, whereby multiple training runs with different start random seeds were carried out to prevent outlier results.

Table 4 shows the language error rates in both language tasks for the two test lengths, measured on the NIST data.

In spite of the relatively poor accuracy of the acoustic models used in separation, adding this information to the phonotactic component consistently improved the overall performance. In particular, considerable improvements were achieved in the nine-language-task where the phonotactics in isolation suffered from the absence of bigrams for the three additional languages within the decoder.

There is a noticeable difference between the two components in the concern of error rates and test lengths: while the robustness of the phonotactic scores increases with longer utterances, the acoustic models seem to exhaust their potential within shorter sequences already, so that the performance does not differ significantly for the two legths.

Final language error rates 9.8%/0.8% and 14.7%/5.6% were reached for 10s/45s utterances in the six- and nine-language-task respectively, which corresponds to improvements by 23%/76% and 35%/40% relative to the phonotactic component (baseline system).

7. DISCUSSION

As expected, the results proved that the performance of an ALI-system can be increased by incorporating multiple information sources. While acoustic modeling in isolation supplied rather insufficient results (being comparable to other work [6]) it can serve with success as a component supporting the phonotactic analysis.

Beside the fact that the described system outperforms comparable systems [2][3], the obvious advantage is a smaller computational expenditure. Whereas in the implicit phonotactic-acoustic modeling [3] the acoustic probabilities are computed at each node in the trellis for several phone-recognizers, in this system just the scores for the recognized phone-sequence are required. Further on, additional languages may be added to the system without the need for manually labeled data.

Future research should address further extension to prosodic features as well as a robust way for language rejection. Also, the question of how to handle non-native speakers represents an open problem. For non-native speakers a divergent behaviour of the two components can be expected. Such a fact might be used for non-nativity detection connected with a proper system response.

8. REFERENCES

[1] Y.K. Muthusamy, E. Barnard and R. A. Cole, "Automatic Language Identification: A Review/Tutorial," IEEE Signal Processing Magazine, October 1994.

[2] Y. Yan, E. Barnard, R. Cole, "Development of an approach to automatic language identification based on phone recognition," Computer Speech and Language, Vol. 10, No. 1, January 1996, pp. 37-54.

[3] M.A. Zissman, "Comparison of four approaches to automatic language identification," IEEE Trans. on Speech and Audio Processing, Vol. 4, No. 1, January 1996, pp. 31-44.

[4] J. Navrátil, W. Zühlke, "Double bigram-decoding in phonotactic language identification," Proc. of ICASSP-97, Munich, Germany, April, 1997, Vol. II, pp. 1115-8.

[5] J. Navrátil, W. Zühlke, "Phonetic-context mapping in language identification," Proc. of EUROSPEECH-97, Rhodes, Greece, September, 1997, Vol. I, pp. 71-4.

[6] T.J. Hazen, V.W. Zue, "Recent improvements in an approach to segment-based automatic language identification,"" Proc. of the 1994 Int. Conf. on Spoken Language Processing, Yokohama, September, 1994, pp. 1883-1886.

[7] Y.K. Muthusamy, R.A. Cole, B.T. Oshika, "The OGI multi-language telephone speech corpus," Proc. of the International Conference on Spoken Language Processing, Banff, Alberta, October, 1992.

[8] T. Lander, R.A. Cole, B.T. Oshika, M. Noel, "The OGI 22 language telephone speech corpus," Proc. of EUROSPEECH-95, Madrid, Spain, September, 1995, Vol. 1, pp. 817-20.