DOUBLE BIGRAM-DECODING IN PHONOTACTIC LANGUAGE IDENTIFICATION

Jiří Navrátil *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

In this paper a phonotactic language identification system that employs a multilingual phone-recognizer with multiple language-dependent grammars to tokenize the spoken signal into several phone-streams is described. For each stream an independent set of language models is used to compute the language scores that are subsequently processed by two classification stages. Thus, the system acquires information from both the original-label and the decoded-phone statistics. A discriminative weighting method is applied in the second stage for better distinguishing between similar languages. A modified language-bigram model, the so-called skip-gram, that allows exploiting of a wider phonotactic context without increasing the estimation costs of a standard bigram, is introduced. Measured on the NIST'95 evaluation set, the described system outperforms the state-of-the-art phonotactic components that use multiple recognizers, and is, at the same time, less computationally expensive.

1. INTRODUCTION

Automatic language identification (ALI) is a task of recognizing the language from a spoken test sentence.

Besides other solutions to this problem based on prosody modeling as well as on phonetic acoustic features [1] there is an efficient way to describe a language in a discriminative way - by means of statistical modeling of phonetic chains (phonotactics). Several contributions have been published dealing with the use of phone n-grams, particularly bigrams, which were shown to be suitable for distinguishing between languages [2],[4].

In the phonotactic components of most ALI systems, one or multiple phonetic recognizers are implemented for tokenizing the incoming utterance in terms of a certain phone repertoire, followed by a set of interpolated bigram language models. Although Zissman and Singer [2] proved that the modeling of phonotactic constraints in terms of the phone set of one language is feasible for identification of several languages, extended systems with multiple language-dependent recognizers were designed in order to better represent the phone repertoire and to improve the overall performance [5],[7].

A somewhat different strategy for applying phonotactic features combined with acoustic modeling was presented in [3] and [7]. Here, multiple language-dependent phone-recognizers processed the utterance, and the resulting acoustic likelihoods were taken for the final classification.

In both cases, the computational costs were considerable due to the multiple recognition process.

In the following, a system will be presented based on a single multilingual phone-recognizer with multiple language-specific bigram models used in the decoding process, followed by stream-dependent sets of language models. This architecture allows a faster phonetic decoding than with the multiple language-dependent recognizers and acquires information of both the original label statistics and the decoded-sequence statistics.

2. ALI SYSTEM WITH DOUBLE BIGRAM-DECODING

The block diagram of the overall identification system is shown in Fig. 1. In the following subsections a description of each of the components is given.

2.1. Phonetic Recognizer

An incoming spoken utterance is decoded by an HMM-based phonetic recognizer with a phone-repertoire broadly covering six languages (55 phones + 6 non-speech units). During the Viterbi decoding, M language-specific bigram (transition) probabilities are applied to constrain the trellis space, which results in obtaining M different phone sequences (streams) on the output of the recognizer. The bigram models that are applied within the recognizer are estimated using manually labelled transcriptions available for the M languages, thus carrying the information about the original-label statistics (later on "inner bigrams").

2.2. Language Models

With every phone-stream from the phonetic recognizer a set of N language-bigram models is connected each exploiting statistical constraints in the corresponding decoded stream for a given language (later on "outer bigrams"). Unlike the inner bigrams, the language models exploit the information of the decoded-phone statistics. Obviously, there are M different language models for each language $L_i \in \{L_1, ..., L_N\}$. As the original properties intristic to a language are changed by the phonetic decoding dependent on the statistical grammar used, the M language models can be told to describe phonotactic properties of the language L_i in terms of the M decoder-inner "languages."

As the core of the language models, standard interpolated bigrams were applied [8] to acquire dependences between

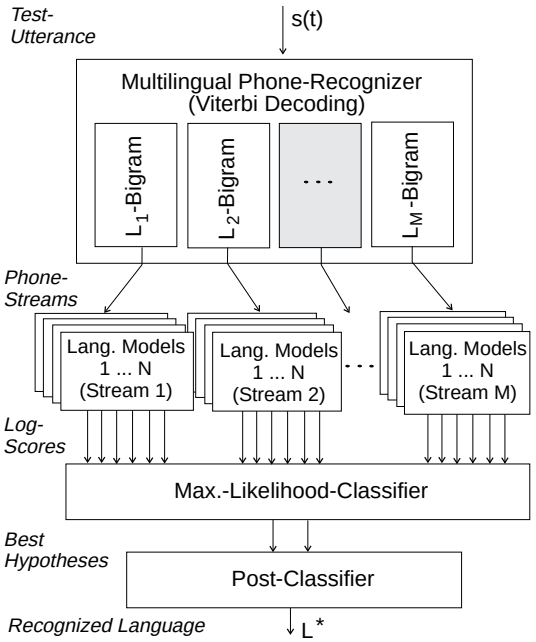

Figure 1. ALI system overview

neighboring phones. Let $A^{(l)} = a_1^{(l)}, ..., a_T^{(l)}$ be the phonetic sequence decoded using the inner bigram l. Then, the phonotatic language score for $A^{(l)}$, given the language L_i, is calculated as:

$$S_{bi}(A^{(l)} \mid L_i) = \tag{1}$$
$$= \frac{1}{T}\left\{ \log \Pr(a_1^{(l)} \mid L_i) + \sum_{t=2}^{T} \log B(a_t^{(l)} \mid a_{t-1}^{(l)}, L_i) \right\}$$

where B denotes the interpolated bigram model:

$$B(a_t \mid a_{t-1}, L_i) = (1-\alpha)\Pr(a_t \mid a_{t-1}, L_i) + \alpha \Pr(a_t \mid L_i)$$

with α the interpolation constant.

It is a well-known fact that statistical dependences are present in a wider context than that of phone pairs in spoken utterances. However, a 2nd-order statistical analysis (trigrams) is faced with the general problem of lacking robustness due to sparse data. In order to overcome this difficulty, and to still capture a wider phonetical context, a sort of modified bigrams were designed as an addition to the core language models. Here, the modified bigram is defined as the a-posteriori probability of a pair of phones not neighboring immediately but with a time gap between them (one phone skipped) as illustrated in Fig. 2.

Later on, such modified bigrams are called skip-grams[1]. Skip-grams can be, similarly to standard bigram, interpreted as a marginal distribution of the joint trigram-probability distribution by summing the time-slot $t-1$ over

[1] It is obvious that the number of phones skipped can be varied which results in different skip-gram models with an even wider context. Our experiments, however, did not show further improvements when using such extended skip-grams.

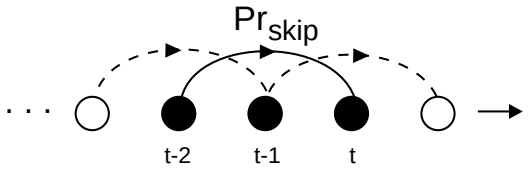

Figure 2. 1st-order analysis with one phone skipped

all phones out of the repertoire $\mathcal{A}$:

$$\begin{aligned}\Pr(a_t \mid a_{t-2}) &= \Pr(a_t \mid *, a_{t-2}) \\ &= \sum_{a_{t-1} \in \mathcal{A}} \Pr(a_{t-2}, a_{t-1}, a_t) / \Pr(a_{t-2})\end{aligned}$$

(Similarly, summing the time-slot $t-2$ would result in obtaining the standard bigram).

Skip-gram combined with the standard bigram can partially exploit context information of a phone-triple in a frame of 1st-order analysis thus not increasing the estimation costs. The interpolated skip-gram score for the test utterance $A^{(l)}$ is calculated in a manner similar to (1):

$$S_{skip}(A^{(l)} \mid L_i) =$$
$$= \frac{1}{T}\left\{ \log \Pr(a_1^{(l)} \mid L_i) + \sum_{t=3}^{T} \log B(a_t^{(l)} \mid a_{t-2}^{(l)}, L_i) \right\}$$

and both the standard bigram and the skip-gram scores can be combined together in an additive way as follows

$$S(A^{(l)} \mid L_i) = (1-\beta) S_{bi}(A^{(l)} \mid L_i) + \beta\, S_{skip}(A^{(l)} \mid L_i)$$

where the influence of the skip-gram can be customized by varying the parameter β.

2.3. Maximum-Likelihood Classifier

For a test speech signal $s(t)$ having been decoded to M streams, M scores for each individual language from the repertoire are computed and put together. The classifier makes a maximum decision based on the total language scores:

$$L^* = \arg \max_{1 \le i \le N} \sum_{l=1}^{M} S(A^{(l)} \mid L_i) \tag{2}$$

Alternatively, the two best hypotheses are determined for later processing in a second stage as described in section 2.4.

2.4. Post-Classification

Optionally, a post-classification method is applied to process two best hypotheses output by the ML-classifier, as introduced in [9]. Here the goal is to rise the significance of individual phone-pairs whose probabilities differ among the languages in the pair given, while suppressing those having similar, i.e. less relevant, bigram values. This is achieved by a discriminative weighting within the scores. As the dissimilarity measure of two languages m and n in the sense of phonotactics a delta-matrix was defined as:

$$\mathbf{\Delta}^{mn} = \{\Delta_{ij}^{mn}\}_{i,j} = \left\{ \frac{\mid B(a_i \mid a_j, L_m) - B(a_i \mid a_j, L_n) \mid}{D_{max}} \right\}_{i,j}$$

(With a norm constant D_{max}) and used to weight the log-probabilities in the score for each of the best languages $L \in \{L_m, L_n\}$:

$$S^*(A^{(l)} \mid L) = \frac{1}{T}\sum_{t=2}^{T} \left(\Delta^{mn}_{a_t,a_{t-1}}\right)^{\gamma} \cdot \log B(a_t \mid a_{t-1}, L_k)$$

(γ is an additional degree of freedom used to control the impact of Δ).

Based on the new scores S^*, the rank-list of best hypotheses may be reordered. Of course, more than two best hypotheses can be taken into account by the post-classifier by processing them in pairs. Extensive experiments have shown, however, that the processing of the two best scores is sufficient for a nine-language-task.

In the framework of the double bigram-decoding system the scores S^* are re-calculated in each stream in isolation first and then added together as in (2) again. If the score sequences computed in the first stage are stored they can easily be used for computing the new weighted scores. In this case, the additional computation costs coming along with the second-stage classification are negligible.

3. DATABASE AND PHONE RECOGNIZER

Up to nine languages from the OGI Multi-Language Telephone Speech Corpus [10] were used for training and system development, and the NIST[2] test set from March '95 involving nine languages was taken for the system evaluations.

Twelve Mel-warped cepstral coefficients, energy as well as their first derivatives, were extracted from the signal waveforms and the cepstral-mean substraction was carried out to suppress channel-dependent feature components.

For the phone-decoder an HMM-based phonetic recognizer was designed by means of the HTK software V2.0. The usual tri-state left-to-right model architecture for each individual HMM applied. 55 selected phonetic plus 6 non-speech HMM's (context-insensitive) were trained on speech signals in six languages, for which manually labelled and segmented transcriptions were available. A total number of 180 "stories-before-tone" (each 45 secons long) served as the data to train the HMM parameters.

For the estimation of the inner grammars, 50 original-label transcriptions in each of the six languages were used. These probabilities weighted the transitions between individual monophone HMM's during the Viterbi decoding.

Further on, 50+10 utterances (45s-stories) in each of the nine languages were decoded by the phonetic recognizer (alternatively with or without applying the inner bigrams) and the resulting sequences served as data for training the outer language models as well as for tuning the system paramaters (β and γ). The data did not overlap with the set used for phonetic training.

The NIST test set (originally recorded in μ-law format) first was converted into linear 16-bit PCM and then processed by the feature extractor described above. The test data in each language consists of 20 45-second phone calls (spontaneous monologues) and ca. 80 10-second excerpts of them as specified in the NIST guideline.

[2] National Institute of Standards and Technology

Configuration	Error Rate	
	10s	45s
Null-Grammar	28.8%	15.0%
Six Inner Bigrams	18.4%	5.0%
+ Skip-Grams	16.3%	5.0%
+ Post-Classifier (2-best)	15.2%	4.2%

Table 1. Error rates on 10/45s utterances in the six-language-task (NIST'95)

Configuration	Error Rate	
	10s	45s
Null-Grammar	38.7%	19.4%
Six Inner Bigrams	27.6%	13.3%
+ Skip-Grams	26.3%	13.3%
+ Post-Classifier (2-best)	25.8%	12.2%

Table 2. Error rates on 10/45s utterances in the nine-language-task (NIST'95)

4. EXPERIMENTS

Performance of the proposed system was tested using a closed set of six and nine languages. In order to assess the efficiency of the double-bigram-decoding method, the conventional null-grammar decoding combined with a single set of language models (e.g. discussed in [2]) was examined as well.

4.1. Six-Language-Task

The following languages were taken for evaluation in the six-language-task: English, German, Hindi, Japanese, Mandarin and Spanish[3]. The same languages also served as the six inner bigram languages due to the availability of their label-transcriptions in the OGI corpus.

Table 1 shows the resulting error rates for the system with six inner bigrams combined with language models consisting of 1) standard bigrams or 2) standard- and skip-grams, and of the second stage post-classifying two competing hypotheses computed in the first stage. For comparison, the performance of a conventional phonotactic component with null-grammar decoding is given also.

A consistent improvement from 28.8% to 18.4% with the 10s-signals and 15% to 5% error rate for the 45s-signals could be achieved with the proposed double-bigram-system. By using the additional skip-grams within the language models, the error rate further decreased with the 10-second utterances. For longer utterances the skip-grams were ineffective which can be explained by a high robustness of the standard bigram scores for such long sequences. Finally, a certain number of hypotheses misclassified in the first stage could be corrected by the post-classifier thus reducing the error rate to 15.2% and 4.2% for 10s- and 45s-long test utterances.

4.2. Nine-Language-Task

Similar performance gains of the individual system configurations can be seen in Table 2 for the the nine-language-task

[3] Also chosen in the NIST evaluations in March '94

in which six languages listed above plus three other languages (French, Tamil and Vietnamese) were involved. In this case, the inner-bigram decoding, together with skip-grams and the post-classifier, brought an improvement from 38.7% to 25.8% and from 19.4% to 12.2% for the 10-second and the 45-second utterances respectively.

5. DISCUSSION

Results obtained in the experiments clearly prove the efficiency of the double-bigram architecture. Measured on the same NIST'95 test set data, the system outperforms the state-of-the-art phonotactic system with six language-dependent recognizers described in [6], and, at the same time, allows a faster, synchronous Viterbi decoding due to the common phone-set. As the phonetic decoding represents the vast part of the overall computation costs this saving is viewed as a considerable advantage. Also, less speech data are necessary for training the single HMM set, as opposed to the full training of six separate recognizers.

As expected, aquiring a wider phonetical context, even though by means of 1st-order statistics only, contributes to a better performance of standard bigrams. Although, for long test sequences the effect of the skip-grams seems to be rather inferior. Moreover, an increased sensitivity to phone-errors could have negatively influenced the modeling potential of skip-grams as the probability of an erroneous deletion or insertion within a triple is higher than that of a phone-pair. Without regard to storage, the skip-grams do not require any additional data for being estimated.

Previously introduced in [9], in connection with a null-grammar decoder, the post-classifier proved to be feasible for the double-bigram system as well, whereby not increasing the computation costs considerably. Because not all language pairs are suitable for the post-classification (fundamentally different languages) the gain of this method was relatively mediocre.

Several attempts to improve the final ML-classification by using a neural network classifier (as introduced in [5]) were undertaken also. Although, with some unique configurations slightly higher identification rates were measured, no consistent results could be obtained in our system. The information contained in the phonotactic scores seems to be separated sufficiently well by the ML-classifier.

In this contribution, the described system has been considered in isolation. Nevertheless, it can be incorporated in a more general ALI system combining phonotactic and prosodic approaches in order to further improve the overall system performance.

REFERENCES

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

[2] M.A. Zissman, E. Singer, "Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling," Proc. ICASSP-94, Adelaide, Australia, April 1994, pp. I-305 - 308.

[3] L.F. Lamel, J.L. Gauvain, "Language identification using phoneme-based acoustic likelihoods," Proc. ICASSP-94, Adelaide, Australia, April 1994, pp. I-293 - 296.

[4] 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.

[5] Y. Yan, E. Barnard, "An approach to automatic language identification based on language-dependent phone recognition," Proc. ICASSP-95, Detroit, MI, USA, 1995.

[6] Y. Yan, "Development of an approach to language identification based on language-dependent phone recognition," Oregon Graduate Institute os Science and Technology, Dissertation October 1995.

[7] 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.

[8] F. Jelinek, "Self-organized language modeling for speech recognition," Readings in Speech Recognition (Waibel, A., Lee, K.-F.) Morgan Kaufman Publishers.

[9] J. Navrátil, W. Zühlke, "Zweistufiges System zur automatischen Sprachen-Identifikation," Proc. of the 7th conference "Elektronische Sprachsignalverarbeitung", Berlin, Germany, November 1996.

[10] 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, Oct. 12-16, 1992.