IMPROVED PHONOTACTIC ANALYSIS IN AUTOMATIC LANGUAGE IDENTIFICATION

Jiří Navrátil
Department of Communication and Measurement, Technical University of Ilmenau
P.O.Box 0565, 98684 Ilmenau, Germany
Tel: +49 3677 69 1145; fax: +49 3677 69 1195
e-mail: jiri.navratil@e-technik.tu-ilmenau.de

ABSTRACT

This paper presents a method for phone-dependent weighting within phonotactic models in automatic language identification. Based on statistical analysis of the phonetic-recognizer behaviour, a phone confidence measure is derived and used to weight the bigram probabilities during testing. The confidence corresponds to the expected decoding stability of individual phones. The proposed method was shown to improve the system performance consistently on a three-language task. The best improvement of the error rate was from 8.4% to 1.8% for the 45-second utterances.

1 INTRODUCTION

Automatic language identification (ALI) is a task of recognizing the language from an unknown spoken test sentence. Besides other solutions to this problem [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 were published dealing with the use of phone bigrams, i.e. the first-order statistics, to exploit statistical constraints of languages. Zissman and Singer [2] used a single English phone recognizer to transcribe the speech to phone sequences and proved that the modeling of phonotactic constraints in terms of the phone set of one language is feasible for identification of all languages. Yan and Barnard [3] employed six combined language-dependent phone recognizers to better represent the wide phone repertoire.

Due to the fact that the language is unknown a-priori, no grammatical decoding is possible during speech recognition, which results in poor accuracy of the phone recognizer. Although, by combining phonotactic models with other approaches, e.g. the prosody modeling, a reasonable system performance can be obtained, the phone transcription accuracy still remains crucial.

In this contribution, an improvement of the phonotactic-based approach to ALI will be presented based on statistical analysis of the transcription behaviour of the recognizer. From this analysis a set of phone confidence parameters can be derived and employed within the bigram models to weight the test phone sequence according to its expected confidence. Because the bigram models are estimated using the decoded sequences, it is not important to exploit the actual phone errors but to describe the transcription stability, i.e. the disambiguity of errors made by the recognizer with respect to a given phone.

The paper is organized as follows: In section 2 a characteristic parameter set is defined and a measure for the phone confidence, as well as the weighting is proposed. Section 3 briefly describes the data used for the experiments whose results are presented in Section 4. Finally, in section 5, some restrictions of the proposed method are stated.

2 PHONE RECOGNIZER ANALYSIS

In order to measure phone accuracy of a recognizer, usually both the original and the decoded sequences are dynamically aligned, and the number of substitutions, deletions and insertions is registered. In addition, a confusion matrix containing the cross-phone substitutions is often generated.

2.1 Error Probabilities

In the following, the recognizer behaviour shall be analysed more comprehensively, taking all three error types into account. For this the aligned phone sequences are analysed statistically by estimating a set of special parameters. The parameters serve as a basis for the confidence measure introduced later in 2.2.

An example of two aligned sequences could look as follows:

```
Original: a - a - b - - - a - c - a - a ...
Decoded : a - - - b - b - a - c - b - a ...
          ---------------------------
Symbolic: a - Da- b - Ib- a - c - Sb- a ...
```

where `Dx`, `Ix` and `Sx` denote deletion, insertion and substitution of the symbol (phone) `x`.

The probability of an input-phone (original label) i being transcribed to a given output-phone j can be estimated as $\Pr_S(i|j) \approx N_{ij}/N_j$, where N_{ij} is the number of occurrences of phone i substituted by j and N_j

is the total number of j observations on output (decoded sequence). The substitution probabilities[1] form a substitution matrix $\overline{\overline{\mathbf{S}}} := \{ \Pr_S(i|j) \}_{i,j}$ whose diagonal elements stand for the correct transcriptions and the out-of-diagonal elements describe the erroneous substitutions.

Regarding deletion errors with respect to an observed output-phone, let a deletion probability be defined as follows:

$$\Pr_D(i|j) = \Pr(\text{“}Phone\ i\ deleted\ previously\text{”} | \text{“}Phone\ j\ on\ output\text{”}) .$$

$\Pr_D(i|j)$ accounts for the deletion frequency of i when appearing prior to j, thus considering the left (first-order) deletion context of j. Note that this probability is conditioned by the occurrence of a deletion, i.e. that $\sum_i \Pr_D(i|j) = 1$.

The insertion probability is formulated similarly:

$$\Pr_I(i|j) = \Pr(\text{“}Phone\ i\ inserted\ previously\text{”} | \text{“}Phone\ j\ on\ output\text{”}) .$$

As with substitutions, both the deletion and insertion probabilities form the D-matrix and I-matrix respectively and all three matrices are stochastic, i.e. their column elements add up to 1 (i-lines, j-columns). These matrices are characteristic for the recognizer $\mathbf{R} = (\overline{\overline{\mathbf{S}}}, \overline{\overline{\mathbf{D}}}, \overline{\overline{\mathbf{I}}})$ and contain the “conclusion” probabilities for error events within an output (decoded) sequence.

2.2 Index of Coincidence

The issue is to determine the “confidence” of a phone observed on output. By confidence, the stability of the recognizer is meant with respect to a given phone. The parameter set $\mathbf{R}$ will be used as a starting-point for computing this confidence. It has to be noted that for the pursued measure the actual transcription errors are of no interest, only the error stability (or disambiguity). A constantly occuring substitution, e.g. “t” $\rightarrow$ “d”, will be exploited by the bigrams during training and will have no longer an effect in the test.

As the measure for confidence the Index of Coincidence (IC) was applied, defined as sum over the squared distribution $\sum_i p_i^2$. The IC can be interpreted as a measure for the roughness of a distribution [4]. Based on the conclusion matrices the IC for an observed phone j is computed as follows

$$IC_F(j) = \sum_i \Pr_F(i|j)^2 \ , \ F \in \{S, D, I\}$$

The value of the IC is maximum when the conclusion distribution is unique (i.e. the transcription is “stable” with respect to the event S, D or I) and minimum when the event is distributed to all phones uniformly. For example, a unique insertion of “eh” constantly appearing prior to “r” would lead to $IC_I(\text{“}r\text{”}) = 1$, i.e. the phone “$r$” is maximum stable with respect to insertions.

2.3 IC-Weighted Identification

The proposed IC measure may be applied directly to weight the phonotactic models. Phones whose transcription is unstable, i.e. who substitute many different input phones and/or are preceded by ambiguous insertions and deletions, shall be suppressed within the score and vice versa. In order to take all three error types into account the individual IC are combined to get a general IC as follows

$$IC(j) = IC_S(j) \Pr(S) + IC_D(j) \Pr(D) + IC_I(j) \Pr(I) \quad (1)$$

whereby $\Pr(F)$ is the occurrence probability of the event F and $\Pr(S) + \Pr(D) + \Pr(I) = 1$.

The value of $IC(j)$ expresses the stability of substitutions leading to j as well as of deletions and insertions preceding the observation j.

For the ALI usually an interpolated bigram log-score of a test sequence $a_1, \ldots, a_T$ for each of the languages is calculated [2]:

$$L = \frac{1}{T} \sum_t \log Bi(a_{t-1}, a_t) \quad (2)$$

where $Bi(a_{t-1}, a_t)$ denotes the interpolated bigram probability.

The confidence-weighted score is proposed as

$$L^* = \frac{1}{T} \sum_t IC(a_t) \cdot \log Bi(a_{t-1}, a_t) \ . \quad (3)$$

Herein the influence of individual phones corresponds to their expected decoding confidence.

3 DATABASE AND PHONE RECOGNIZER

Three languages out of the Oregon Graduate Institute Telephone Speech Corpus [5] were used to evaluate the modified phonotactic models. For this, the “story-before-the-tone” (story-bt) utterances (45 seconds long) were taken in two similar languages (English, German) and one more different language (Tamil). Further extension to the six-language-task as well as the nine-language-task is planned connected with the future use of a multi-language phone recognizer (see discussion 5).

The speech parameters energy, mel-weighted cepstral and delta cepstral features were computed and the cepstral mean substraction was carried out. The commercial HMM software package HTK 2.0 was used to implement a single English phone recognizer. 40 monophone HMMs consisting of three emitting states and conventional parameter configuration were trained on English-labeled data which did not overlap the English ALI set.

[1] Later on in this paper, substitutions are understood as a general phone-to-phone transform, i.e. inclusive of the correct transcriptions.

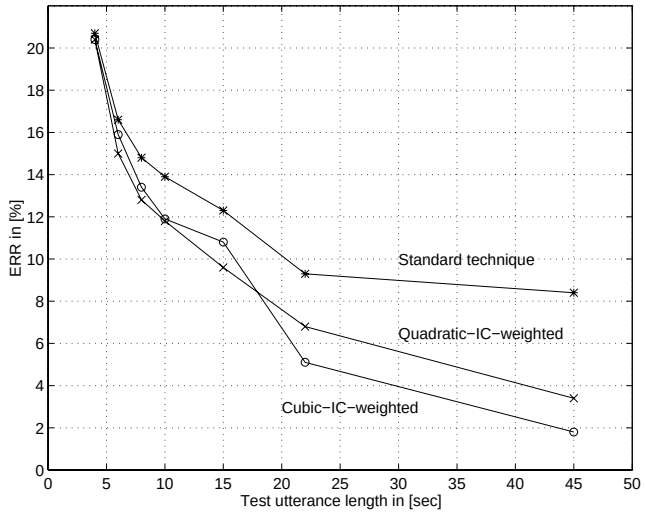

Figure 1: ALI error rates for the three-language-task

The recognizer performed with 40.6% phone accuracy on English test data, which is comparable to other published results [2].

To train the phonotactic bigram models 50 story-bt-sentences decoded by the English recognizer in each language were available. The ALI system performance was evaluated with another 20 story-bt-sentences per language, additionally cut into shorter sequences in order to investigate the trade-off between performance and utterance length.

4 EXPERIMENTS

The recognizer parameters $(\overline{\overline{\mathbf{S}}}, \overline{\overline{\mathbf{D}}}, \overline{\overline{\mathbf{I}}})$ were estimated analysing the aligned original and decoded sentences of the English test data (non-overlapping the ALI set). Then the general IC for each phone were computed, whereby the relative frequences of the individual event types were 0.81, 0.15 and 0.04 for substitutions, deletions and insertions respectively.

Additionally, a modified "Cubic-IC" was proposed for the tests. This IC computes the sum over the cubic distribution instead of the squared one, thus strengthening the element coincidence:

$$IC_F(j) = \sum_i Pr_F(i\,|j)^3 \quad , F \in \{S, D, I\}\ .$$

The ALI performance was evaluated on test utterances with varying length ranging from 4 to 45 seconds. The results for the standard interpolated bigram score (2) and both the quadratic and the cubic IC-weighted scores (3) are shown in Fig. 1.

It can be seen that the IC-weighting consistently improves the system performance and that the gain increases as the test utterance becomes longer. This can be explained by a higher score robustness when longer sequences are processed so that the confidence weighting is more effective.

Error Rate			
Standard	IC(S)	IC(S+D)	IC(S+D+I)
13.9%	12.3%	11.8%	11.8%

Table 1: ERR on 10-second utterances when including separate events

For longer utterances (roughly over 20 seconds) the cubic IC seems to be better suitable than the quadratic IC whereas for shorter tests the cubic IC lacks score robustness and performs slightly worse than the quadratic one. The best improvement was reached for 45-second utterances where the identification error rate decreased from 8.4% to 1.8%.

Table 1 shows the particular error rates for special (quadratic) IC-combinations measured on 10-second utterances. Including only substitutions in the IC, in (1), the error rate improves from 13.9% to 12.3%. By adding the IC_D an improvement to 11.8% could be achieved whereas the addition of IC_I did not further decrease the error rate. This may be explained by the low occurrence frequency of insertions $\Pr(I) \approx 0.04\%$. With longer utterances (over 20 seconds) all combinations performed identically. In that case, the substitutions alone were sufficient to exploit the gain potential of the weighting technique.

5 CONCLUSION

This paper presented a weighting method for phonotactic language modeling which, based on statistical phone recognizer analysis, modifies the bigram language score according to the transcription stability (confidence) of individual phones. There are some theoretical and practical limitations for the parameters derived in this paper:

- only first-order transcription context was considered, i.e. the substitutions depended on one input phone and one deletion or insertion preceding the output phone. In reality this context might be higher. However, a higher-order error analysis meets with the problem of insufficient robustness.
- all parameters are based on English phone vocabulary. Although it was shown this can be used for several languages a more precise acquisition may be expected by introducing a wider vocabulary covering phones of all the languages.

The proposed method does not increase computational costs essentially. Once the transcription analysis is completed all weights can be precomputed and saved for the classification.

Further investigations are planned including a multilingual phone recognizer to cover the phone repertoire of several languages. Future evaluations will be extended to a six-language task as well as to a nine-language-task

as specified by the National Institute of Standard and Technology.

6 ACKNOWLEDGEMENTS

This work is supported by the Thuringian Graduate Grant. The author wishes to thank Professor W. Zühlke for supervising the project and the Center for Spoken Language Understanding at the Oregon Graduate Institute for making available the OGI-ML corpus as well as other information and tools to this database.

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", ICASSP-94, Adelaide, Australia, April 1994, pp. I-305 - 308.

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

[4] W.F. Friedman, "The index of coincidence and its applications in cryptanalysis", Technical Paper, 1925, (Available through Aegean Park Press, Laguna Hills CA).

[5] 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.