# Improved *N*-gram Phonotactic Models For Language Recognition

*Mohamed Faouzi BenZeghiba, Jean-Luc Gauvain and Lori Lamel*

Spoken Language Processing Group
LIMSI - CNRS B.P. 133 91403 ORSAY CEDEX FRANCE

## Abstract

This paper investigates various techniques to improve the estimation of *n*-gram phonotactic models for language recognition using single-best phone transcriptions and phone lattices. More precisely, we first report on the impact of the so-called *acoustic scale factor* on the system accuracy when using lattice-based training, and then we report on the use of $n$-gram cutoff and entropy pruning techniques. Several system configurations are explored, such as the use of context-independent and context-dependent phone models, the use of single-best phone hypotheses versus phone lattices, and the use of various $n$-gram orders. Experiments are conducted using the LRE 2007 evaluation data and the results are reported using the *a posteriori* EER. The results show that the impact of these techniques on the system accuracy is highly dependent on the training conditions and that careful optimization can lead to performance improvements.

**Index Terms**: Language recognition, Phonotactic model, Phone lattices

10.21437/Interspeech.2010-718

## 1. Introduction

Phonotactic approaches for language recognition rely on the assumption that the way phones are arranged in words and sentences is language specific [1]. This means that even if two languages have the same set of phonemes, their phonotactic characteristics are different. Phonotactic approaches try to capture these differences and to use them to discriminate between languages.

In practice, phonotactic constraints are automatically inferred from the speech signal. The inference requires the availability of a well trained speech recognizer. This recognizer is used to decode speech signals producing phone sequences or phone lattices from which phone *n*-gram statistics are estimated and an *n*-gram phonotactic model is built.

Work on phonotactic approaches has mainly focused on improving the quality of phone recognizer with the goal of making the recognizer more accurate and the phone *n*-gram counts more reliable. Several techniques have been investigated such as using more training data [2], phone lattice decoding [3], better phone modeling [4] and CMLLR adaptation [4, 5]. All these techniques have significantly improved language recognition performances, in particular the use of phone lattices.

Surprisingly, little work has focused on improving the estimation of phonotactic model from the phone *n*-gram counts. This paper investigates two ways of improving the phonotactic models, one based on optimizing the *acoustic scale factor* with lattice-based model training and the other uses an *n*-gram cutoff to eliminate infrequent and presumably incorrect *n*-grams. For comparison purposes, entropy pruning is also investigated. While the use of these techniques in speech recognition is a common practice, their impact on the phonotactic language recognition system have not yet been investigated. The paper is organized as follows. The next section presents the main principles of the language recognition problem. Then is followed by a description of how the phonotactic models are estimated, and the interplay between the phone lattices and the acoustic scale factor, and the application of *n*-gram cutoffs during model estimation. Experimental results are given in Section 4 showing how these techniques influence the language recognition performance.

*This work has been partially supported by OSEO, French State agency for innovation, under the Quaero program.

## 2. Phonotactic approach

For language recognition using phonotatic models, the decision is made by maximizing the following language score:

$$S(X, L) = \log \sum_{H} f(X|H, L, \theta) P(H|L) \qquad (1)$$

where $f(X|H, L, \theta)$ is the likelihood of the speech segment $X$ given a phone sequence $H$, a language $L$ and the acoustic model $\theta$. The probability $P(H|L)$ is the phone sequence probability given by the language-dependent phonotactic model. This score can be approximated by considering only the most likely phone sequence $H^\star$:

$$S(X, H^\star, L) = \log f(X|H^\star, L, \theta) + \log P(H^\star|L) \qquad (2)$$

In addition, under the assumption that the acoustic model is language independent, the speech segment likelihood $f(X|H^\star, L, \theta)$ can be replaced by $f(X|H^\star, \theta)$ which is a constant across all languages and does not affect the decision. Therefore assuming that we have an *n*-gram phonotatic model, the language score can be simplified as follows:

$$S(X, H^\star, L) \simeq \sum_{h_1^n} C(h_1^n) \log P(h_n|h_1^{n-1}, L) \qquad (3)$$

where $C(h_1^n)$ is the frequency of the phone *n*-gram $h_1^n$ in the hypothesis $H^\star$ and where the summation is taken over all observed *n*-grams in $H^\star$.

If a phone lattice $\mathcal{L}_X$ is used instead of keeping only the most likely phone sequence, equation (3) still holds if the *n*-gram frequencies are replaced by the expectation of the *n*-gram frequencies given the phone lattice as follows:

$$S(X, H^\star, L, \mathcal{L}_X) = \sum_{h_1^n} E[C(h_1^n)|\mathcal{L}_X] \log P(h_n|h_1^{n-1}, L) \qquad (4)$$

where the summation is taken over all *n*-grams in the phone lattice.

One of the main problem is to get accurate estimate of the n-gram probabilities $P(h_n|h_1^{n-1}, L)$ for each targeted language.

# 3. Phonotactic model estimation

With 1-best phone decoding, the maximum likelihood estimates of the phonotactic model probabilities $P(h_n|h_1^{n-1}, L)$ are obtained by counting the number of times the phone *n*-gram $h_1^n$ occurred and dividing the count by the number of occurrences of the context $h_1^{(n-1)}$. Several smoothing techniques are proposed to get better estimates. For language recognition, Witten-Bell smoothing was found to perform the best [8]. In this work, an interpolated version of the Witten-Bell algorithm is used [9].

## 3.1. Phone lattices and acoustic scale factor

Phone lattices are graphs where nodes correspond to particular speech frames and where edges represent the phone hypotheses and have associated acoustic scores. Phone lattices are generated by a phone decoder using context-independent or context-dependant acoustic models without any phonotactic constraints. The *n*-gram probabilities are estimated by taking the expected frequencies given the phone lattice $\mathcal{L}$[3]:

$$E[C(h_1^n) \mid X, \theta] \simeq \sum_{h(e_i)=h_i} P(e_1, ..., e_N|\mathcal{L}) \qquad (5)$$

where in the right hand part we compute the sum of the lattice posterior probabilities of all sequences of $N$ edges corresponding to the phone *n*-gram $(h_1, ..., h_N)$. The lattice posterior probabilities in (5) are computed by means of the forward-backward algorithm which gives us:

$$P(e_1, ..., e_N|\mathcal{L}) = \alpha(e_1)\beta(e_N) \prod_i \xi(e_i)/\beta_0 \qquad (6)$$

where $\alpha(e)$ is the forward likelihood of the starting node of the edge $e$, $\beta(e)$ is the backward likelihood of the ending node of edge $e$, $\beta_0$ is the backward likelihood of the first vertex, and $\xi(e)$ is the likelihood of the edge $e$ estimated as follows:

$$\xi(e) = \delta f(X_e|\theta_e)^{1/\gamma} \qquad (7)$$

where $f(X_e|\theta_e)$ is the likelihood of the speech segment $X_e$ corresponding the lattice edge $e$ given the HMM phone model associated to the edge. The parameter $\gamma$ is the acoustic scaling factor used to compensate for the HMM independency assumptions and for the model size. The parameter $\delta$ is the phone insertion penalty corresponding to a uniform phone language model.

Even though $\gamma$ can be seen as the inverse of the well known language model weight, for lattice decoding the distinction is quite important. This acoustic scaling factor is also commonly used in speech recognition when using word lattices, in particular for MMIE training [6], consensus decoding [7], and to get accurate confidence scores from lattices. Without this scaling factor the best hypothesis incorrectly dominates the alternate solutions, making the estimation of the posterior probabilities unreliable. The role of the factor $\gamma$ is to control the distribution of the posterior over phones [10]. In this work we do not report on tuning the parameter $\delta$ as we found that such tuning has little impact on language recognition performance. It was set at a value experimentally found to give good performance.

## 3.2. $N$-gram cutoff and entropy pruning

Applying *n*-gram cutoffs is a well known technique in language modeling that consists of excluding from the model those *n*-grams that occur less than a certain number of times (typically one or two) in the training data. These low count *n*-grams are often considered to be irrelevant, but since they account for a large part of the observed *n*-grams, including them increases the memory requirements. The main purpose of applying count cutoffs in speech recognition, is to reduce considerably the language model size without significantly affecting its accuracy.

In language recognition, the problem is more serious, since phone *n*-grams are estimated from the output of the recognizer and not from text or manual speech transcripts. Building a phone recognizer that can capture fine phonotactic characteristics for several languages is a challenging task, due to the mismatch in the (acoustic) characteristics between the language used to train the recognizer and the target languages. This mismatch can considerably affect the accuracy of the recognizer and make the results of the decoding rather approximate. In this case it is likely that low-count phone *n*-grams are not good indicators of the target language and can introduce noise which degrades the discriminability of the models. This is particularly true with less accurate systems, for example when decoding with context-independent phone models and using just the 1-best phone sequence rather than phone lattices to estimate the phonotactic models. As will be seen in the results section, using a count cutoff is found to be efficient for improving language recognition performance for certain system configurations, in particular when 1-best phone sequence decoding is used.

For comparison purposes, entropy pruning, which consists of reducing the number of phone *n*-grams by a certain percentage, is also applied.

# 4. Experiments and Results

## 4.1. Experimental conditions

Language recognition systems explored in this work make use of the PPRLM (Parallel Phone Recognizer followed by Language Model) with 3 phone recognizers, one for English, Spanish and French. Both context-independent (CI) and context-dependent (CD) phone models are explored. A more detailed description of these systems can be found in [4] with a few minor modifications: the English recognizer has 38 phones instead of 48 and CD models cover about 2000 context rather than 3000. We find that these new systems run much faster without affecting language recognition performances.

The language training data are selected from several sources[1] (LRE-96 train and Dev. NIST LRE'07 train, Callhome, Mixer and Ficher databases) and performance is evaluated using the NIST LRE'07 eval data, containing 14 languages[2]. The acoustic feature vector has 39 dimensions (12 PLP + energy + $\Delta$ + $\Delta\Delta$). Phonotactic models with different orders (*2*-gram, *3*-gram and *4*-gram) are generated from both 1-best hypothesis and phone lattices.

The language detection decision is made based on the average posterior probability estimated from language likelihoods and target and non-target language *priors* using Bayes rule. For better analysis of the results, performances are reported in terms of *a posteriori Equal Error Rate* (EER).

[1] Defined by MIT Lincoln Labs.

[2] http://www.nist.gov/speech/tests/lang/2007/LRE07EvalPlan-v8b.pdf

### 4.2. Impact of the acoustic scale factor

In this section, all phonotactic models are generated from phone lattices without cutoffs or entropy pruning. Figures (1) and (2) plot the EER as a function of the *scaling factor* for different phonotactic model orders using CI and CD phone models, respectively. Optimal scaling factors with their corresponding EER are reported in Table (1) for both CI and CD phone models.

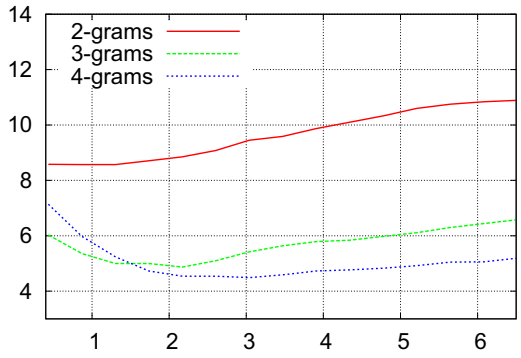


Figure 1: EER[%] as a function of the scaling factor using phone lattices generated with CI phone models.

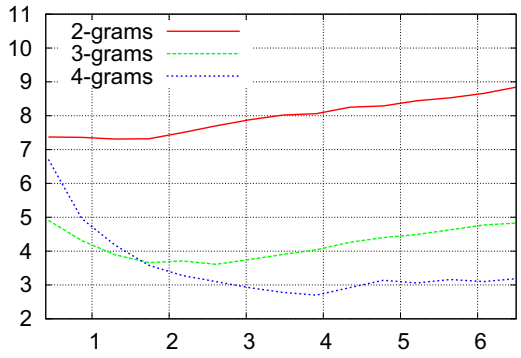


Figure 2: EER[%] as a function of the scaling factor using phone lattices generated with CD phone models.

| *Models* | | *2-gram* | *3-gram* | *4-gram* |
|---|---|---|---|---|
| CI | EER[%] | 8.6 | 4.9 | 4.5 |
| | $\gamma$ | 1.1 | 2.2 | 3.0 |
| CD | ERR[%] | 7.3 | 3.6 | 2.7 |
| | $\gamma$ | 1.3 | 2.6 | 3.9 |

Table 1: EER with the optimal acoustic scale factor ($\gamma$) using CI and CD phone models.

From the two figures, it is clear that optimizing the scaling factor with phone lattices can significantly improve language recognition performance. The gain is dependent upon the phonotactic model order and the acoustic model type. It can be seen that while there is almost no effect of the scaling factor with the *2*-gram models, it is more important with the *3*-gram and *4*-gram models. One possible explanation is that since the 3 vocabularies associated with phone recognizers are small[3], using the few best hypotheses only (corresponding to $\gamma = 0.4$), is sufficient to accurately estimate *2*-gram statistics from several hours of speech. Indeed all *2*-grams have high expected counts. In contrast, to reliably estimate the expected counts of the higher order *n*-grams, phone lattices need to be explored deeply to extract as much information as possible. The contribution of the alternative phone hypotheses becomes more important and the value of the scaling factor needs to be increased. High order *n*-grams are more language-specific and therefore more discriminant, estimating them reliably leads to a significant improvement. The superiority in terms of consistency of CD phone models over CI phone models was demonstrated and largely discussed in a previous work [4].

It can be observed also that, the value of the scaling factor is not so high. In fact, The scaling factor controls the sharpness of the posterior probability distribution. Without this factor, the distribution will be very sharp, reducing the generalization capabilities of the phonotactic models. However, a high value will result in a very flat distribution. In this case, phonotactic models for different languages will have almost the same distribution, reducing their discriminant capabilities. Therefore, the value of the scaling factor is a trade-off between the generalization and the discriminant capabilities of the estimated models.

Finally, the optimal values of the scaling factor obtained with CD models are generally higher compared to their corresponding values obtained with CI models. This can be explained by the difference in the model size between CI and CD models.

### 4.3. Effect of count cutoffs and entropy pruning

#### *4.3.1. Using 1-best hypothesis*

Table (2) reports on the results of using entropy pruning and count cutoffs for *n*-grams generated from 1-best hypothesis. The parameters of the cutoff model $(N_2, N_3, ..., N_i)$ corresponding to the minimum count for all possible *n*-gram orders in the phonotactic models are given in brackets. All unigrams are kept in the phonotactic model ($N_1 = 1$).

| *Models* | *Conditions* | *2-gram* | *3-gram* | *4-gram* |
|---|---|---|---|---|
| CI | no cutoff | **8.7** | 7.9 | 18.1 |
| | cutoffs | (1) | $(1, 4)$ | $(1, 2, 4)$ |
| | | **8.7** | **5.2** | **5.2** |
| | entropy pruning | 10.1 | 6.7 | 6.3 |
| CD | no cutoff | **7.4** | 5.7 | 14.0 |
| | cutoffs | (1) | $(1, 4)$ | $(1, 3, 4)$ |
| | | **7.4** | **4.2** | **3.7** |
| | entropy pruning | 8.3 | 5.6 | 4.9 |

Table 2: Effect of the *n*-gram cutoffs and entropy pruning on the EER for CI and CD phone models using 1-best hypothesis.

It can be seen that while the best results for *2*-grams are obtained without cutoffs (for the same reasons explained in Section (4.2)), for *3*-gram and *4*-gram models, significant improvements are obtained using count cutoffs for CI and CD models. With CI models, the relative improvements are 28% and 71% for the *3*-gram and *4*-gram models, respectively. With CD models, these improvements are 26% and 74%, respectively. The

[3]The size of English, French and Spanish vocabularies are 38, 36 and 27 phones, respectively.

relatively high values of the cutoff parameters associated with *3*-gram and *4*-gram (e.g. using CD models, phone *3*-grams and *4*-grams with counts less than 3 and 4, respectively, are not included in the *4*-gram phonotactic models) indicate that most of these *n*-grams are not estimated reliably. As explained in Section (3.2) and unlike speech recognition where these irrelevant *n*-grams do not affect system performance, in language recognition they can be considered as noise. It is interesting to note here, that for *4*-gram phonotactic model, in particular those generated with CI models, the irrelevant phone *n*-grams represent most of the model size.

More interestingly, comparing results in Table (2) to those plotted in Figures (1) and (2), language recognition system using 1-best phone hypothesis with count cutoffs technique performs significantly better than those using phone lattices without scaling factor optimized (except for *2*-gram model where performances are equivalent).

Improvements with entropy pruning of the phonotactic models are not as significant as those obtained with count cutoffs. The entropy pruning technique applied in this work consists of reducing the size of the model by a certain percentage. This implies removing those *n*-grams that less affect the perplexity of the model. While these technique works relatively well with *4*-gram, it degrades performances with *2*-gram models.

#### *4.3.2. Using phone lattices*

We have applied count cutoffs and entropy pruning techniques to phone *n*-gram counts estimated from phone lattices with different values of $\gamma$. Table (3) reports the results.

| *Models* | *Conditions* | *2-gram* | *3-gram* | *4-gram* |
|---|---|---|---|---|
| CI | no cutoff | 8.6 | 4.9 | 4.5 |
| | cutoffs | (0) | (0, 0) | (0, 0, 0) |
| | $\gamma$ | 1.1 | 2.2 | 3.0 |
| | | 8.6 | 4.9 | 4.5 |
| | entropy pruning | 9.6 | 4.9 | 4.5 |
| CD | no cutoff | 7.3 | 3.6 | 2.7 |
| | cutoffs | (0) | (0, 0) | (0, 0, 0) |
| | $\gamma$ | 1.2 | 2.6 | 3.9 |
| | | 7.3 | 3.6 | 2.7 |
| | entropy pruning | 8.3 | 3.7 | 2.8 |

Table 3: Effect of the *n*-gram cutoffs and entropy pruning on the EER for CI and CD phone models using phone lattices. The optimal values of the cutoffs and scaling factor are given model order.

Surprisingly, the best performances are obtained without cutoffs[4]. This means that all *n*-grams are important and removing some of them degrades system performance. For a possible explanation, it is worth remembering that the expected *n*-gram counts estimated from phone lattices are based on *posteriori probabilities* which are a more reliable measure than using *n*-gram frequencies. As $\gamma$ is increased to its optimal value, more sources (in terms of number of phone sequences) are used. This makes the expected *n*-gram counts more precise and more accurate. Even those with low counts are language specific and contribute to the quality of the phonotactic model.

[4]Since the *n*-gram counts generated from phone lattices are float, we have to take into account counts less than one.

The same observations can be made for entropy pruning. The reported results are usually obtained by reducing the model size with only $1\%$.

## 5. Conclusion

This paper has explored the optimization of the acoustic scale factor in the phone lattice decoding framework, and the use of count cutoff and entropy pruning to reduce the phone $n$-gram model size, with the aim of improving the accuracy of the phonotactic language recognizer. The importance of optimizing the acoustic scale factor with lattice-based decoding was demonstrated. Concerning LM pruning, applying low count cutoffs significantly improved the language recognition performance with 1-best decoding, while entropy pruning was not very effective. Neither pruning method was found to be effective with lattice-based decoding, showing that phone lattices provide reliable phone n-gram estimates for phonotactic language recognition.

## REFERENCES


[1] M. P. Harper and M. Maxwell, "Spoken Language Characterization" Springer Handbook of Speech Processing, Chap. 40, pp. 797-807, 2008

[2] P. Matejka, P. Shwarz, J. Cernocky and P. Chytil, "Phonotactic Language Identification using High Quality Phoneme Recognition", *Proceedings of Eurospeech'05*

[3] J. L. Gauvain, A. Messaoudi and H. Schwenk, "Language Recognition Using Phone Lattices", *ICSLP'04*

[4] M.F. BenZeghiba, J.L. Gauvain and L. Lamel, "Context-Dependent Phone models and Models Adaptation for Phonotactic Language Recognition", *Interspeech'08*, pp. 313-316.

[5] W. Shen and D. Reynolds, "Improved Phonotactic language recognition with acoustic adaptation" *Proc. of Interspeech'07*

[6] P. C. Woodland and D. Povey, "Large Scale Discriminative Training for Speech Recognition" *ASR2000* Automatic Speech Recognition Challenges for the new Millenium.

[7] L. Mangu, E. Brill and A. Stolcke, "Finding Consensus Among Words: Lattice-Based Word Error Minimization" *Proc. Eurospeech'99*

[8] L. Wang, E. Ambikairajah and E. H. C. Choi, "Multilingual Phoneme Recognition and Language Identification Using Phonotactic Information" *Proc. of ICPR'06*, Vol.4, pp.245-248,

[9] S. F. Chen and J. Goodman, "An empirical Study of Smoothing Techniques for Language Modeling" *Compt. Speech Lang.*, Vol 13, pp. 359-393, 1999.

[10] W. M. Campbell, F. Richardson and D. A. Reynolds "Language Recognition With Word Lattices And Support Vectore Machines" *ICASSP 2007*