Multi-Stream Statistical N-Gram Modeling With Application To Automatic Language Identification

Katrin Kirchhoff, Sonia Parandekar

Department of Electrical Engineering
University of Washington, Seattle, USA
{katrin,sonia}@ee.washington.edu

Abstract

Most state-of-the art automatic language identification systems are based on phonotactic information, i.e. languages are identified on the basis of probabilities of phone sequences extracted from the acoustic signal. This approach ignores the potential advantages to be gained from a richer representation of the acoustic signal in terms of parallel streams of subphonemic events. In this paper we develop an alternative approach to language identification which is based on parallel streams of phonetic features and sparse modeling of statistical dependencies between these streams. We present results on the OGI-TS database and show that the feature-based system outperforms a comparable phone-based system significantly while using fewer parameters. Moreover, the feature-based system exhibits a markedly better performance on very short test signals ($\leq$ 3 seconds). The theoretical approach developed here is of significance not only for language identification but also for related work in pronunciation modeling.

1. Introduction

A recent trend in speech technology is the rapidly growing need for multilingual speech processing systems. Automatic language identification (LID) is an integral part of such systems and therefore continues to attract much attention from the research community. One of the most important problems, particularly for embedded speech applications, is the development of algorithms for fast language identification using very short test signals. Most current LID systems continue to use the phonotactic approach developed over the past decade (e.g.[7, 8, 9]). Under this approach, sequences of phones are extracted from the speech signal using standard acoustic front ends, such as hidden Markov models (HMMs). These phone sequences are then rescored by language-specific statistical n-gram models. Language-discriminating information is thus assumed to be encoded primarily in the statistical patterns phone sequences in different languages. One of the drawbacks of this approach is the problem of unseen phone contexts in the test data, e.g. when applying a system trained on read speech to spontaneous speech [10]. Moreover, phone-based systems need a fairly large n-gram context for accurate LID, which is why a they deteriorate rapidly on short test signals (less than 10 seconds).

More powerful modeling schemes are conceivable, in particular models which employ multiple sequences of subphonemic events such as articulatory or phonetic features like *vocalic*, *nasal*, *voiced*, etc. In the field of automatic speech recognition, such models have recently regained attention [1, 2, 3, 4]. As we will describe in greater detail below, a multiple phonetic feature stream approach might also be of benefit for language identification. In particular, improvements may be expected with respect to the amount of training material needed to train or to adapt a LID system to new languages, the length of the acoustic test signal, memory requirements and accuracy. In the following section we will describe the basic theoretical model of the new approach and contrast it with the standard phone-based model. In Section 3 we describe the corpus used for the present study and our baseline phone-based and feature-based LID systems. Section 4 discusses experiments on feature-stream selection and score combination. In Section 5 we address the issue of modeling statistical dependencies between different streams of phonetic features. Experimental results and conclusions are are given in Sections 6 and 7, respectively.

2. A Feature-Based Approach to Language Identification

The standard phone-based approach to LID assumes that most of the language-discriminating information is contained in the statistical regularities governing phone sequences in different languages. To model this effect, an acoustic front end, typically consisting of phone HMMs, is used to map the acoustic signal to a sequence of phone symbols. This front end can be language-independent, i.e. trained on data from all languages in the system, or language-dependent, in which case different acoustic models are trained for each language. Throughout this paper we will use language-independent acoustic models for both the phone-based and the feature-based LID systems. The model sequences obtained through recognition using the acoustic models are then used to train language-specific n-gram models. For language identification, all n-gram models are applied to a given test sequence and the language corresponding to n-gram model that assigns the highest probability to the sequence is assumed to be the true language. This decision rule can be described formally as

$$L^* = argmax_L P(\phi_1, \phi_2, ..., \phi_N|L) \tag{1}$$

where L^* is the index of the best hypothesized language, and $P(\phi_1, \phi_2, ..., \phi_N|L)$ is the probability of the phone sequence $\phi_1, \phi_2, ..., \phi_N$ given language L, which is computed according to the phone n-gram model for that language:

$$P(\phi_1, \phi_2, ..., \phi_N|L) = \prod_{i=n}^{N} P(\phi_i|\phi_{i-1}, ..., \phi_{i-n+1}) \tag{2}$$

Typically, smoothed bigrams or trigrams are used as n-gram models.

Under the new feature-based approach multiple parallel sequences of phonetic features are extracted from the signal. Phonetic features are abstract classes loosely related to articulatory properties, such as *voiced*, *vowel*, *nasal*, *rounded*, etc. They can be grouped into subsets depending on their possibilities of co-occurrence: all features which mutually exclude each other are assigned to the same group. Thus, features describing *voicing* form a subset, as do features describing *manner* of articulation, *place* of articulation, etc. These subsets implicitly identify partially independent dimensions of articulation.

For each feature a statistical acoustic model is built, analogous to acoustic phone models. Acoustic decoding is carried out for each feature group independently, yielding parallel phonetic feature streams. For each feature stream, an individual feature n-gram model is estimated. During testing, the preprocessed signal is thus passed through a bank of feature recognizers, followed by feature n-grams which compute the probabilities of the each feature stream given the language. The resulting stream-specific scores are combined to provide the overall LID score. Analogous to the n-gram probability of a phone sequence, the probability of a feature stream $\mathcal{F} = f_1, ...f_N$ given a language L, $P(\mathcal{F}|L)$, is defined as the product of the conditional probabilities of the current feature given the n-gram context:

$$P(\mathcal{F}|L) = \prod_{i=n}^{N} P(f_i|f_{i-1}, ..., f_{i-n+1}) \tag{3}$$

The probability of an ensemble of K feature sequences $\mathcal{F}_1, ..., \mathcal{F}_K$ given language L, $P(\mathcal{F}_1, ..., \mathcal{F}_K)$, is

$$P(\mathcal{F}_1, ..., \mathcal{F}_K|L) = \mathcal{C}(P(\mathcal{F}_1|L), , ..., P(\mathcal{F}_K|L)) \tag{4}$$

where $\mathcal{C}$ is some combination function. The best language hypothesis is then identified by the maximum score:

$$L^* = argmax_L P(\mathcal{F}_1, ..., \mathcal{F}_K|L) \tag{5}$$

A simple combination function, which we use in our baseline feature-based system, is the product of the individidual feature stream scores:

$$P(\mathcal{F}_1, ..., \mathcal{F}_K|L) = \prod_{k=1}^{K} P(\mathcal{F}_k|L) \tag{6}$$

It should be noted that this model might have limitations in that all feature streams are assumed to be independent given the language. We will address the problem of incorporating cross-stream dependency modeling below.

The feature-based approach has a range of potential advantages: First, the number of feature classes in any given language is typically much smaller than the number of phone classes; in fact, approximately 30 phonetic features suffice to encode all phones in the world's languages. Training data for phonetic features can be shared across phones, leading to a larger number of training samples per class. Feature n-gram models and acoustic models can therefore be trained more robustly than the corresponding phone models. Moreover, the overall number of models (and therefore the number of parameters in the system) is reduced.

Second, the language-independent nature of phonetic features enhances the portability of a feature-based LID system to new languages or dialects. Whereas phone-based systems often struggle with the problem of unseen phones or phone sequences when confronted with new data, the potential range of unseen feature contexts is much smaller.

Third, a large part of cross-linguisitic variation arises from differences in articulatory timing, e.g. different degrees of vowel nasalization or aspirated vs. unaspirated plosives. Such phenomena may have language-differentiating potential. In a phone-based system, however, they can only be modelled at the expense of creating additional phone models (e.g. for nasalized vs. non-nasalized vowels), thereby enlarging the number of parameters to be trained. This type of variation can be modelled more adequately when phones are represented as sets of phonetic features: characteristic shifts in articulatory timing can be expressed as changes to the statistical dependencies between features in different streams, without having to enlarge the model inventory. A further, related advantage is that shorter test signals might be needed to for accurate language identification since subphonemic contexts can be exploited more thoroughly.

3. Corpus and Baseline Systems

For the experiments described in the present study we use the OGI-TS corpus [5] of multilingual telephone speech. The training, development and evaluation set definitions are the ones established in [6]. The sets contain 4650, 1898 and 1848 files for training, development and evaluation, respectively, with approximately the same number of files for each of ten different languages (English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese). It should be pointed out that this data differs from most OGI-TS subsets used for previous studies on language identification in that very short signals ($\leq$ 3 seconds) are included, whereas most previous evaluation set definitions (e.g. those used for the NIST LID evaluations in 1993-1995) have excluded signals shorter than 10s. We expressly intended to include very short signals in order to test our hypothesis that the feature-based approach might be beneficial in situations where little acoustic test material is available.

We built two baseline LID systems, one based on phones, the other based on phonetic feature streams. The phonetic feature groups we use are *voicing*, *manner* of articulation, *consonantal place* of articulation, *nasality*, *vocalic place* of articulation, *front-back* tongue-position, and lip *rounding.* The phone set comprises 126 models, the total number of feature models is 47. Furthermore, 7 models are used for silence, short pauses, and various types of noise, laughter, etc. Both systems were bootstrapped on manual phone labels for a small subset of the data (623 files); phone labels were converted to phonetic feature labels using the definitions of the International Phonetic Alphabet (IPA). All acoustic feature and phone models are single-Gaussian HMMs with 3 states. After bootstrapping, automatic transcriptions were created by unconstrained phone and feature recognition passes, i.e. no top-down information in the form of language models or phone/feature grammars was used to constrain recognition.

These automatic transcriptions were used to train phone and feature back-off n-gram models. The results obtained by the baseline systems on forced-choice identification of the 10 languages listed above, are listed in Table 1 for different n-gram contexts. All feature streams in the feature-based system had the same n-gram context.

Our baseline systems use acoustic models with a uniform topology and thus do not take into account characteristic variations in duration of certain phones and features. To overcome this limitation we investigated a simple form of durational modeling, where the number of states in each acoustic model was adjusted to reflect the average duration of the model's instances in the training data. The number of states was set to the average

	n=2	n=3	n=4	n=5	n=6
phone	47.2	42.3	38.8	38.3	37.9
feature	37.6	43.9	44.7	42.4	40.6

Table 1: Baseline LID accuracies (in %) of phone-based and feature-based systems, development set.

duration divided by the acoustic frame rate of the system (10 ms). This form of duration modeling increased the accuracy to 50.3% for the phone-based system and 46.7% for the feature-based system.

Note that the current systems do not use clustered models, speaker adaptation or multiple Gaussian mixtures - the global performance level of both systems could certainly be improved if these features were added. We plan to integrate these over time; in our initial experiments, priority was given to exploring the modeling options of the new feature-based approach.

4. Feature Stream Selection and Score Combination

In LID experiments using only a single feature stream at a time it was observed that different feature groups vary greatly with respect to their language-discriminating potential. Consonantal place features are most informative whereas voicing and nasality features seem to be the least useful. However, the combination of the feature streams with the best individual results does not necessarily lead to the best overall result due to nonlinear interactions between the streams. For this reason we selected the best subset of feature streams by optimizing the LID accuracy on the development set. The best subset achieves a LID accuracy of 48.8% on the development set and consists of the five streams *manner*, *consonantal place*, *vowel place*, *front-back*, and *rounding.*

As a more advance score combination scheme we used a Multi-Layer-Perceptron to map vectors of normalized feature stream scores for all languages to the final LID probabilities. Additional experiments were performed using the feature stream rank instead of normalized scores. However, neither of these schemes led to any any improvement over simple product combination.

5. Cross-Stream Dependency Modeling

One of the weaknesses of the baseline feature-based system is that cross-stream dependencies are not taken into account. To improve the accuracy of the system it seems to be crucial to indentify relevant dependencies to incorporate them into the statistical model.

In the baseline model, the probability of a feature f in stream j at position i, $P(f_i^j)$, is only dependent on the previous $n-1$ features in the same stream. When cross-stream dependencies are incorporated, it is additionally dependent on a set of feature in streams other than j:

$$P(f_i^j) = P(f_i^j | f_{i-1}^j, ..., f_{i-n+1}^j, \mathcal{F} \setminus \{j\}) \quad (7)$$

If we assume a topological ordering on the streams, the set $\mathcal{F} \setminus \{j\}$ can contain any number of features at any position up to i in any stream less than than j. For our initial experiments, however, we simplified this model by allowing only pairwise dependencies between two streams and only between features which overlap in time. If we furthermore assume that all streams are independent given j, we can make the following approximation:

$$\begin{aligned} P(f_i^j) &= P(f_i^j | f_{i-1}^j, ..., f_{i-n+1}^j, \mathcal{F} \setminus \{j\}) && (8) \\ &\approx \prod_{k=1, k \prec j}^{K} P(f_i^j | f_{i-1}^j, ..., f_{i-n+1}^j, f_i^k) && (9) \end{aligned}$$

The additional conditioning variable f_i^k can be considered part of the n-gram context, in which case the number of potentially observable contexts increases from m_j^n to $m_j^n \times m_k$ (where m_j and m_k are the number of different feature values in streams j and k, respectively). In order to robustly estimate probabilities for these events, and in particular for unseen contexts, the same smoothing and back-off procedures used in standard n-gram modeling can be applied. However, we can also approximate the above quantity as

$$\begin{aligned} P(f_i^j) &= P(f_i^j | f_{t-i}^j, ..., f_{t-n+1}^j, f_i^k) && (10) \\ &\approx P(f_i^j | f_{t-i}^j, ..., f_{t-n+1}^j) P(f_i^j | f_i^k) && (11) \end{aligned}$$

which has the advantage that fewer parameters, namely $m_j^n + m^j \times m^k$, need to be estimated.

Another key question is how to identify those cross-stream dependencies which are designed to improve LID accuracy, i.e. which have the strongest discriminating effect. This concep is called structural discriminability [11]. In principle, dependencies can be found either by a search through the space of possible dependencies, or by using various heuristics to predict the effect of adding individual dependencies. To test the first option we have used the greedy search algorithm described below;

Greedy Dependency Search

Initialization
instantiate list of dependencies to all possible pairwise combinations of feature streams
set i=0
set S_0^{max} to LID score for model without dependencies
while list of dependencies not empty
do
    set $i = i + 1$
    for all dependencies $d \epsilon D$ in list
    do
        add dependency d
        compute new LID score, S_i^d
        set $\delta_i^d = S_i^d - S_{i-1}^{max}$
        remove dependency d
    done
    find dependency d with the largest δ_i^d
    if $\delta_i^d > 0$ and d does not create circular dependency
        add d to model
        remove d from list
    else if $\delta_i^d < 0$
        terminate
    set $S_i^{max} = S_i^d$
done

The constraint on circular dependencies ensures that the result is a valid probability distribution.

To test the second option, we use the mutual information between two streams X and Y as a selection criterion:

$$I(X;Y) = \sum_{x,y} P(x,y) log \frac{p(x,y)}{p(x)p(y)} \quad (12)$$

Dependencies between streams showing the highest mutual information are added until the LID accuracy on the development decreases, again observing the circular dependency constraint.

6. Experiments

We first applied the greedy search strategy to our best feature-based LID system (i.e. the 5-stream model described in Section 4); the LID accuracy was optimized on the development test. Dependencies were implemented as conditional probability tables; the integration of the cross-stream probabilities was done according to the model in Equation 11. The results are shown in Table 2. We notice that the LID accuracy increases markedly

Iteration	Dependency	LID acc (%)
0	no dependencies	48.8
1	cons. place - rounding	53.6
2	manner - vowel place	54.6
3	front/back - cons. place	55.1

Table 2: LID accuracy on development test after adding cross-stream dependencies. Dependencies are listed as "dependent variable" - "conditioning variable".

already after the first iteration. Further added dependencies contribute smaller improvements. The total improvement over the best result obtained by the phone-based system (50.3%) is statistically significant at the 0.002 level. We then tested the dependency selection based on mutual information. This procedure terminated after two iterations with a development test accuracy of 53.8%. The dependencies identified by this criterion were *front-back - vowel place* and *front-back - consonantal place*. Finally, we applied our best phone-based system, the best feature-based system without dependencies and the best feature-based system with dependencies to the evaluation set. Results are listed in Table 3.

System	LID acc (%)
phone	50.7
feature without dependencies	49.2
feature with dependencies	56.7

Table 3: LID accuracy of best systems on evaluation set.

In order to compare the performance of the different systems on test signals of varying lengths we grouped the evaluation set files into three categories, viz. very short (shorter than 3s), short (between 3s and 15s), and long (longer than 15s), and scored these separately. Results are shown in Table 4. We see

System	very short	short	long
phone	33.3	54.8	70.9
feature without deps	40.2	50.8	62.9
feature with deps	48.0	58.8	64.6

Table 4: LID accuracy (in %) of best systems on evaluation files of different lengths.

that the feature-based system achieves a much higher performance on very short test signals; this characteristic is further enhance by cross-stream dependency modeling.

7. Summary and Conclusions

In this paper we have developed a novel approach to language identification which is based on n-gram models of parallel streams of phonetic features and sparse statistical dependencies between these streams. We have shown that such a multi-stream feature-based system outperforms a comparable phone-based system significantly while using fewer parameters. Moreover, the feature-based system shows a particularly large improvement on very short test signal (≤ 3 seconds). So far we have only explored a small number of the possible modeling options which this new approach offers. Future work will include exploring further data-driven measures for predicting optimal cross-stream dependencies, as well as different schemes for score integration.

Acknowledgements
The work presented in this paper was funded by US DoD. We would like to thank the Oregon Graduate Institute for providing us with the manual phone labels for the OGI-TS data.

8. References

[1] K. Kirchhoff, "Combining acoustic and articulatory information for speech recognition in noisy and reverberant environments", *Proceedings ICSLP-98*, pp. 891–894, 1998

[2] Kirchhoff, K., G.A. Fink and G. Sagerer, "Conversational Speech recognition using acoustic and articulatory input", *Proceedings of ICASSP-00*, pp. 1435–1438, 2000

[3] King, S. and P. Taylor, "Detection of phonological features in continuous speech using neural networks", *Computer, Speech and Language 14(4)*, pp. 333–353, 2000

[4] M. Ostendorf, "Incorporating linguistic theories of phonological variation into speech recognition models," *Royal Society Phil. Trans.*, 2000, to appear

[5] Muthusamy, Y.K., R.A. Cole and B.T. Oshika, "The OGI multi-language telephone speech corpus", *Proceedings of ICSLP-92*, 1992

[6] Muthusamy, Y.K., *A Segmental Approach to Automatic Language Identification*, PhD Thesis, Oregon Graduate Institute, October 1993

[7] Zissman, M.A., "Comparison of four approaches to automatic language identification of telephone speech", *IEEE Trans. Speech and Audio Processing 4(1)*, pp.31–44, 1996

[8] Yan, Y. and E. Barnard, "Experiments for an approach to language identification with conversational telephone speech", *Proceeedings ICASSP-95*, pp. 789–792, 1995

[9] Corredor-Ardoy, C. et al., "Language identification with language-independent acoustic models", *Proceedings Eurospeech-97*, pp. 55-59, 1997

[10] Zissman, M.A., "Predicting, diagnosing and improving automatic language identification performance", *Proceedings Eurospeech-97*, pp. 51–54, 1997

[11] Bilmes, J., "Dynamic Bayesian Networks", *Proceedings of 16th Conference on Uncertainty in Artificial Intelligence*, 2000