Hierarchical Multilayer Perceptron based Language Identification

David Imseng[1,2], *Mathew Magimai.-Doss*[1], *Hervé Bourlard*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland
{dimseng,mathew,bourlard}@idiap.ch

Abstract

Automatic language identification (LID) systems generally exploit acoustic knowledge, possibly enriched by explicit language specific phonotactic or lexical constraints. This paper investigates a new LID approach based on hierarchical multilayer perceptron (MLP) classifiers, where the first layer is a "universal phoneme set MLP classifier". The resulting (multilingual) phoneme posterior sequence is fed into a second MLP taking a larger temporal context into account. The second MLP can learn/exploit implicitly different types of patterns/information such as confusion between phonemes and/or phonotactics for LID. We investigate the viability of the proposed approach by comparing it against two standard approaches which use phonotactic and lexical constraints with the universal phoneme set MLP classifier as emission probability estimator. On SpeechDat(II) datasets of five European languages, the proposed approach yields significantly better performance compared to the two standard approaches.

Index Terms: Language identification, multilingual processing, hierarchical MLP.

1. Introduction

The goal of automatic language identification (LID) is to classify a given input speech utterance as belonging to one out of N languages. Various possible applications of LID can be found in multilingual speech processing, call routing and interactive voice response applications.

There are a variety of cues, including phonological, morphological, syntactical or prosodic cues, that can be exploited by an LID system [1]. In literature, different approaches have been proposed to perform LID, such as using only low level spectral information [2], using phoneme recognizers in conjunction with phonotactic constraints [3, 4] or using medium to high level information (e.g. lexical constraints, language models) through speech recognition [5]. Among these, the most common approach is to use phoneme recognizers along with phonotactic constraints. The phoneme recognizer can be language-dependent [4] (using a language specific phoneme set) or it can be language-independent [6] (using a multilingual phoneme set). The phonotactic constraints are typically modeled by a phoneme bigram estimated on phonetically labeled data.

In this paper, we propose a hierarchical MLP-based approach for language identification. The proposed approach tries to model information, such as confusion among phonemes and phonotactics present in long temporal sequences ($\approx$ 150-300 ms) of phoneme posterior probabilities. We demonstrate the viability of the proposed approach using five European languages from the SpeechDat(II) corpus.

The remainder of this paper is organized as follows. In Section 2, we present the motivation for the proposed approach. Section 3 describes the used database and Section 4 briefly describes the investigated systems. Section 5 discusses the experimental results and Section 6 concludes the paper.

2. Motivation

The hierarchical MLP-based approach for language identification that is proposed in this paper is inspired by a recently proposed hierarchical MLP-based approach for phoneme posterior estimation [7] [8] .

In the hierarchical MLP-based phoneme posterior estimation approach, first an MLP is trained to classify phonemes in a conventional manner using standard cepstral features as input. A second MLP is then trained to classify phonemes but with the phoneme posterior probabilities (posterior features) estimated from the first MLP with a temporal context of around 150-230 ms as input feature. On phoneme recognition tasks as well as speech recognition tasks, it has been found that the hierarchical approach yields a better performance compared to conventional single MLP-based approaches [8]. Upon analysis of the second MLP using Volterra series, it was found that the second MLP learns phonetic-temporal patterns present in the posterior features. The learned phonetic-temporal patterns consist of acoustic confusions among phonemes and phonotactic constraints of the language [8].

In the context of language identification, such phonetic-temporal patterns could possibly be exploited by first training an MLP to classify a "universal" phoneme set (multilingual speech units), and then modeling the resulting posterior features (with a long temporal context) by a second MLP to classify languages. It can be expected, that information related to phonotactic constraints and acoustic confusion among phonemes (present in the posterior features spanning a long temporal context) is language specific.

The motivation behind using a universal phoneme set is that it allows data sharing and discriminant training between phonemes across languages. Furthermore it can help in bootstrapping systems for unseen languages [9].

3. Database

We use data from SpeechDat(II) that currently consists of recordings from 14 different European countries. In order to be representative, the SpeechDat(II) databases are gender-balanced, dialect-balanced according to the dialect distribution in a language region and age-balanced. The databases are subdivided into different corpora. For our preliminary study, we used *Corpus A*, that contains three read application words per speaker. The term *application words* describes a set of about

30 words such as “help” or “cancel”, which could be used in interactive voice response applications.

In the presented work, the datasets of five languages, namely British English (EN), Swiss French (SF), Swiss German (SZ), Italian (IT), and Spanish (ES) were used. In Swiss German, there are 2000 recorded speakers. As standardized by SpeechDat(II), datasets with a minimum of 2000 speakers have pre-defined test sets that contain the data of 500 speakers. The remaining 1500 speakers are sub-divided into a development set (10%, 150 speakers) and a training set (1350 speakers). To avoid any bias in terms of available amount of data towards a particular language, the same number of speakers was used in all languages, even if other databases provide data from more than 2000 different speakers. For this purpose, a subset of 2000 speakers was chosen from the whole dataset by using the same procedure as for the test set creation and then the subset was split into training, development, and test set. Hence, we did not use the pre-defined test sets, but rather used the scripts available at [10] to ensure that the splits can be reproduced.

Table 1 gives information about the data of each language, including the number of utterances, the mean duration of the utterances and the minimal utterance duration (after voice activity detection).

Table 1: *Statistics of the datasets. The number of utterances that are available for each language as well as mean and minimal duration of the utterances are displayed.*

Language	utterances		duration	
	total	testset	mean	min
English (EN)	5207	1305	1.20 s	0.31 s
Spanish (ES)	5817	1447	1.23 s	0.31 s
Italian (IT)	5416	1368	1.53 s	0.31 s
Swiss French (SF)	5668	1429	1.34 s	0.32 s
Swiss German (SZ)	5720	1426	1.21 s	0.32 s

We use the lexicon provided along with the database. The lexicon contains word pronunciations in terms of the SAMPA[1] phoneme set. Table 2 displays the number of phonemes that are used to transcribe the application words of different languages. Note that some languages do not use all the available phonemes for the application words task.

Table 2: *Number of phonemes used per language for the application words task.*

Language	EN	ES	IT	SF	SZ
# phonemes	33	29	35	36	46

In order to create a universal phoneme set, we merged the phonemes that share the same SAMPA symbol across languages. In Table 3, the poly-phonemes which are used by more than one language are displayed and it is shown by how many languages a particular poly-phoneme is shared. For each language, the remaining mono-phonemes are also given. As seen in Table 4, the Italian and the Swiss German databases have the most mono-phonemes in their dictionaries. Table 4 also displays the phoneme sharing factor of all the languages that shows by how many languages the phonemes of a particular language are shared on average. The Spanish phonemes for instance are shared by 3.3 language on average.

Table 3: *Universal SAMPA phoneme set with all the poly- and mono-phonemes. Silence is shared across all languages, thus the universal phoneme set consists of 92 phonemes.*

Poly-phonemes (37)	
Shared by	phonetic symbols
5 lang.	d, k, l, n, s, t, g, f, p, m
4 lang.	j, e, v, b, a
3 lang.	@, r, S, w, i, u
2 lang.	tS, dZ, I, u:, i:, aI, N, h, R, x, E, o, J, z, 9, O
Mono-phonemes (54)	
Language	phonetic symbols
EN	{, O:, eI, Q, I@, @U, 3:
ES	jj, D, rr, T, B, L, G
IT	’u, ’o, nn, ll, ’a, ’E, ’i, SS, ddz, mm, ’e, ttS, ss
SF	A, O/, a∼, &/, y, o∼, Z, e∼, H
SZ	?, U, aU, 2:6, a:, OY, 2:, ts, y:, e:, o:, E:, C, i:6, Y, E6, o:6, U6
Silence	sil (shared by all languages)

Table 4: *The number of mono-phonemes per language and the phoneme sharing factor for all languages.*

Language	EN	ES	IT	SF	SZ
# of mono-phonemes	7	7	13	9	18
phoneme sharing factor	3.1	3.3	2.9	3.1	2.5

4. System Description

All the approaches studied here use an MLP trained to classify a universal phoneme set consisting of 92 phonemes. As shown in Fig. 1, the input to the MLP is nine frames of 39 dimensional perceptual linear prediction (PLP) cepstral coefficients consisting of 13 static coefficients (including zeroth), their approximate first and second derivatives. The PLP features were extracted at a frame rate of 10 ms with a frame size of 25 ms after having performed voice activity detection using Tracter[2]. We refer to this MLP as phoneMLP.

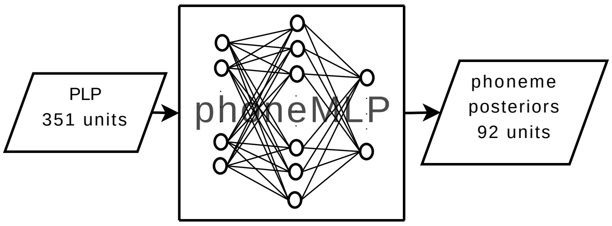

Figure 1: *Illustration of the universal phoneme set classifier. The MLP is referred to as phoneMLP.*

4.1. LID using Phonotactic Constraints (PC)

The phonotactic constraint based approach exploits low-level knowledge i.e., phonemes and phoneme sequences for language identification. We denote the system based on phonotactic constraints as *System PC*.

In System PC, a test utterance is processed by five parallel language-specific HMM/MLP [11] phoneme recognizers.

[1] http://www.phon.ucl.ac.uk/home/sampa/index.html

[2] http://juicer.amiproject.org/tracter/

Each phoneme recognizer consists of a fully connected ergodic model [4] connecting all the 92 phoneme HMMs (each phoneme is modeled with a three state left-to-right HMM). A phoneme bigram language model models only the phoneme transitions allowed in the pronunciations of the words corresponding to the language. In this study, the words are the application words corresponding to each language. The phonotactic constraints/phoneme bigram models are obtained from the respective lexicon. The emission likelihoods of the HMM states are estimated from the output of the phoneMLP. The language corresponding to the phoneme recognizer output that yields the highest likelihood score is picked as the recognized language. Figure 2 illustrates the System PC, where the parallel systems correspond to the language-specific phoneme recognizers.

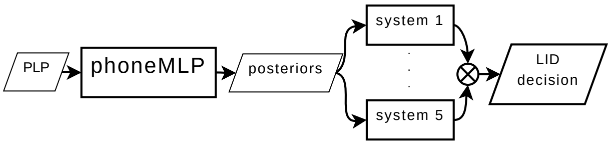

Figure 2: *Using a different system for each language. The system yielding the highest score is identified as the language.*

4.2. LID through Speech Recognition (SR)

The approach of performing LID through speech recognition tends to exploit higher level prior knowledge such as, lexicons and language models/syntactical constraints. We denote the system corresponding to this approach as *System SR*.

In System SR, a test utterance is processed by five parallel hybrid HMM/MLP speech recognizers (in this study, isolated word recognizers) one corresponding to each language. The dictionaries contain all the test words (no out-of-vocabulary words). Each phoneme is modeled with a three state left-to-right HMM and the emission likelihoods of the HMM states are estimated from the output of the phoneMLP. The language corresponding to the speech recognizer that yields the word hypothesis with maximum likelihood is chosen as the recognized language. Figure 2 illustrates the System SR as well, where the parallel systems now correspond to the isolated word recognizers of different languages.

4.3. Hierarchical MLP-based LID (Hier)

We denote the system based on the hierarchical MLP-based approach proposed in this paper as *System Hier*. Figure 3 gives a schematic view of the System Hier. In this system, an MLP (referred to as LID-MLP) is trained to classify languages using the phoneme posteriors estimated by the phoneMLP as input feature. We vary the temporal context at the input of the LID-MLP and study its impact on the performance of the LID system. When varying the temporal context, the number of hidden units is accordingly adjusted to keep the number of parameters constant. Given a test utterance, the frame-based $\log$ posteriors for each language are summed up and the decision about the language is made by choosing the language that gets the maximum $\log$ posterior probability over the whole utterance.

In retrospect, it can be observed that the different systems described in this section use the output of the phoneMLP differently. More specifically, System PC and System SR use the phoneMLP output as local score (acoustic match) and try to discriminate between languages using lower level or higher level "a priori" knowledge (i.e. knowledge driven). However, the

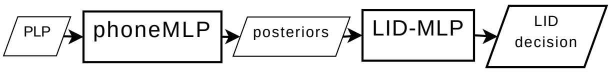

Figure 3: *The hierarchical approach. The "phoneMLP" is shown in Fig. 1 and the "LID-MLP" is sketched in Fig. 4.*

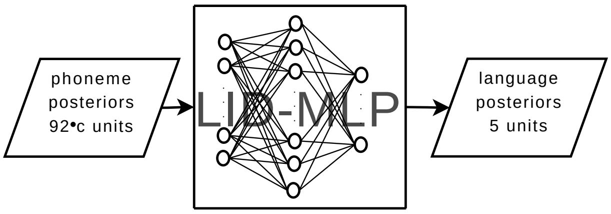

Figure 4: *Architecture of the LID-MLP. The input dimensionality depends upon the temporal context (c frames) which is varied in this study. At the output are five units, one for each language.*

System Hier uses the output of phoneMLP as a feature, and learns in a data driven manner to discriminate between languages.

5. Experimental Results and Discussion

We performed language identification on the test set of the five SpeechDat(II) datasets for English, Spanish, Italian, Swiss French and Swiss German. In total there are 6975 available test utterances. The System Hier was evaluated for different temporal contexts at the input of the second MLP (LID-MLP). The temporal context was varied from one frame (10 ms) up to 310 ms (minimal utterance duration). Table 5 presents the performance of different systems.

Table 5: *Comparison of different systems. The System Hier performance was obtained with a temporal context of 290 ms at the input of the LID-MLP.*

System	Errors	LID %
PC	1236	82.3
SR	360	94.8
Hier	248	**96.4**

The results show that System Hier (with 290 ms temporal context) yields a significantly better performance (McNemar with 99% confidence level) compared to both, System SR and System PC. Figure 5 presents the influence of the temporal context on the performance of the hierarchical MLP-based approach. It can be observed that an increasing temporal context improves the language classification accuracy and saturates at a temporal context of around 230 ms. This trend is similar to what has been observed in the case of hierarchical MLP-based phoneme recognition. It can also be seen that System Hier improves over System SR at a temporal context of around 130 ms or above. Furthermore, it is interesting to notice that with no temporal context (where one may expect only acoustic confusion related information to be present), the hierarchical MLP-based approach yields a better performance than the phonotactic constraint-based approach.

In order to better understand the difference between System Hier and System SR, we analyzed the confusion between different languages. Tables 6 and 7 display the confusion be-

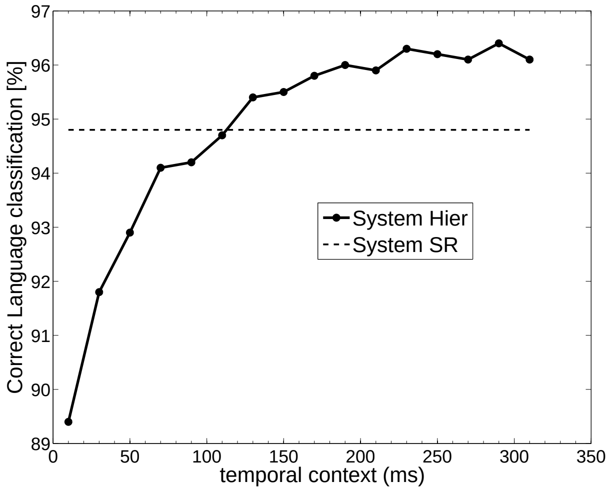

Figure 5: *Influence of the temporal context to System Hier. The performance of System SR is significantly worse compared to System Hier with a temporal context ≥ 170 ms.*

tween different languages for System Hier and System SR, respectively. False negatives represent the number of misclassifications per language. The false negatives are also given as percentage of the total amount of test utterances available for a particular language. False positives on the other hand, indicate how many times a particular language was wrongly associated to a test utterance of another language.

Table 6: *Confusion between languages for System Hier (290 ms temporal context).*

	EN	ES	IT	SF	SZ	false neg.	
EN	-	9	23	5	10	47	3.6%
ES	6	-	32	6	11	55	3.8%
IT	4	18	-	4	7	33	2.4%
SF	1	7	12	-	50	70	4.9%
SZ	5	2	18	18	-	43	3.0%
false pos.	16	36	85	33	78	248	

Table 7: *Confusion between languages for System SR.*

	EN	ES	IT	SF	SZ	false neg.	
EN	-	30	24	10	27	91	7.0%
ES	5	-	15	2	2	24	1.7%
IT	6	53	-	6	2	67	4.9%
SF	14	27	7	-	57	105	7.3%
SZ	25	13	8	27	-	73	5.1%
false pos.	50	123	54	45	88	360	

The misclassification rates are more even across languages in System Hier than in System SR. In the case of System Hier, the languages Italian and Swiss German yield low misclassification rates but at the same time have more false positives. This may be due to the fact that these languages have a high number of mono-phonemes (see Table 4). In the case of System SR, the Spanish language yields the lowest misclassification rate but at the same time higher false positives. This may be attributed to the nature of the Spanish mono-phonemes and the high phoneme sharing factor (see Table 4). English and Swiss French also have a high sharing factor, but their mono-phonemes contain mostly vowel sounds, whereas the Spanish mono-phonemes are rather consonant sounds.

Altogether, the findings of our study suggest that there is a good potential in using the proposed hierarchical MLP-based approach for language identification.

6. Conclusion and Future Work

In this paper, a hierarchical MLP-based approach that tries to model phonetic-temporal patterns in phoneme posterior sequences was proposed for language identification. Experimental studies that used SpeechDat(II) databases of five languages demonstrated that the proposed approach can yield a system that performs significantly better than systems based on conventional approaches that use phoneme recognition with phonotactic constraints or a speech recognition system.

In future, we intend to further ascertain the potential of the proposed approach by using more languages, continuous speech data, and using other techniques proposed in the literature to create a universal phoneme set.

7. Acknowledgments

This research was supported by the Swiss NSF through the project MultiModal Interaction and Multimedia Data Mining under contract number MULTI-200020-122062 and through the National Center of Competence in Research on “Interactive Multimodal Information Management” (www.im2.ch). The authors would like to thank Phil Garner for his help with Tracter.

8. References

[1] M. A. Zissman and K. M. Berkling, “Automatic language identification,” *Speech Communication*, vol. 35, pp. 115–124, 2001.

[2] M. Sugiyama, “Automatic language recognition using acoustic features,” in *Proc. of ICASSP*, 1991, pp. 813–816.

[3] J. Navratil, “Spoken language recognition - a step toward multilinguality in speech processing,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 9, no. 6, pp. 678–685, 2001.

[4] L. Lamel and J.-L. Gauvain, “Cross-lingual experiments with phone recognition,” in *Proc. of ICASSP*, vol. 2, 1993, pp. 507–510.

[5] T. Schultz, I. Rogina, and A. Waibel, “LVCSR-based language identification,” in *Proc. of ICASSP*, vol. 2, 1996, pp. 781–784.

[6] K. Berkling and E. Barnard, “Theoretical error prediction for a language identification system using optimal phoneme clustering,” in *Proc. of Eurospeech*, 1995, pp. 351–354.

[7] J. Pinto, B. Yegnanarayana, H. Hermansky, and M. Magimai-Doss, “Exploiting contextual information for improved phoneme recognition,” in *Proc. of ICASSP*, 2008, pp. 4449–4452.

[8] J. Pinto, G. S. V. S. Sivaram, M. Magimai.-Doss, H. Hermansky, and H. Bourlard, “Analysis of MLP based hierarchical phoneme posterior probability estimator,” *to appear in IEEE Trans. on Audio, Speech, and Language Processing*, 2010.

[9] B. Wheatley, K. Kondo, W. Anderson, and Y. Muthuswamy, “An evaluation of cross-language adaptation for rapid HMM development in a new language,” in *Proc. of ICASSP*, 1994, pp. 237–240.

[10] G. Chollet *et al.*, “LE2-4001 Deliverable Identification,” ENST, Telenor, CPK and CSELT, Tech. Rep., 1998.

[11] N. Morgan and H. Bourlard, “Continuous speech recognition: An introduction to the hybrid HMM/connectionist approach,” *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 24–42, May 1995.