ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Phonotactic Language Recognition Using MLP Features

Mohamed Faouzi BenZeghiba, Jean-Luc Gauvain and Lori Lamel

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

10.21437/Interspeech.2012-545

Abstract

This paper describes a very efficient Parallel Phone Recognizers followed by Language Modeling (PPRLM) system in terms of both performance and processing speed. The system uses context-independent phone recognizers trained on MLP features concatenated with the conventional PLP and pitch features. MLP features have several interesting properties that make them suitable for speech processing, in particular the temporal context provided to the MLP inputs and the discriminative criterion used to learn the MLP parameters. Results of preliminary experiments conducted on the NIST LRE 2005 for the closed-set task show significant improvements obtained by the proposed system compared with a PPRLM system using context-independent phone models trained on PLP features. Moreover, the proposed system performs as well as a PPRLM system using context-dependent phone models, while running 6 times faster.

Index Terms: Language recognition, Phonotactic approach, MLP features

1. Introduction

Phonotactic language recognition approaches have been shown to benefit from the use of well-known speech recognition techniques such as context-dependent (CD) phone models [1], CMLLR adaptation [1] [2], and phone lattices [3]. Such context-dependent phonotactic systems [4], [5] obtain results that are close to those other more complex state-of-the-art systems. However, CD systems are computationally more expensive than context-independent (CI) systems. For the LIMSI system[1] submitted to the NIST LRE11 evaluation, the average processing speed was 3xRT, making them not well suited for most real applications. Another well-known drawback of phonotactic approaches (compared to acoustic approaches) is the degradation in performance on short speech segments.

Multilayer Perceptron (MLP) derived features, which are growing in popularity for speech recognition, are investigated to address these two issues. These features have two interesting properties essential for building efficient phonotactic-based language recognition systems: long-span temporal context and discriminability. Capturing the context at the feature level (e.g. the Shifted Delta Cepstra [6], [7]) or the model level (e.g. CD phone model or Artificial Neural Network [8]) has been shown to significantly improve system performance. Second, MLP training focuses on the phonetic content while reducing non relevant information. Features derived from the MLP network are therefore more accurate than conventional PLP features, and phone recognizer trained with these features are more consistent. Using such recognizers, phone lattices with better quality can be generated and more accurate phonotactic models can be built.

Two kinds of MLP features can be distinguished, *probabilistic* features extracted from the outputs (representing posterior probabilities of acoustic classes) of the MLP and *bottle-neck* features extracted from the hidden layer of the MLP. In this work, experiments are conducted using bottle-neck features concatenated with conventional PLP and pitch features. The importance of the pitch feature in the concatenated mlpplpf0 features for language recognition task was not investigated yet. However, these concatenated features were found to perform best on speech recognition tasks such as in [9] for Mandarin and for other languages discribed in internal Quaero reports. [2]

To the best to our knowledge, the only prior work investigating MLP features for language recognition task is described in [5]. In that work, PPRLM system performances were improved by using a multilingual phone recognizer trained with the concatenated *probabilistic* MLP and PLP features. As is shown in Section 3.2, our results suggest that better improvement can be achieved by using several recognizers trained with the concatenated features, suggesting that the individual PRLMs provide more complementary information than PLP-based PRLMs. Finally, this paper provides an analysis of the behavior of the PPRLM systems and their PRLM components based on the concatenated feature vectors and when CMLLR adaptation is applied.

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

[1] This system uses three CD phone models, CMLLR adaptation and phone lattices

[2] www.quaero.org.

2. PPRLM System Descriptions

All evaluated systems are based on the PPRLM phonotactic approach, with three phone recognizers for English (EN), Spanish (SP) and French (FR), and vary the type of acoustic features and phone models (CI or CD).

Two types of acoustic features are used. The first are PLP-like [10] features and second are bottleneck features produced by Multi Layer Perceptron (MLP) [14]. Previous experiments with alternate MLP features have shown that the TRAP-DCT features [11] have comparable performance to the warped linear predictive temporal patterns but are much cheaper to obtain.

For the PLP features, 39 cepstral parameters are derived from a Mel frequency spectrum, with cepstral mean removal and variance normalization carried out on a segment-cluster basis, resulting in a zero mean and unity variance for each cepstral coefficient [12]. The TRAP-DCT features are obtained from a 19-band Bark scale spectrogram, using a 30 ms window and a 10 ms offset. As in [13] a discrete cosine transform (DCT) is applied to each band (the first 25 DCT coefficients are kept) resulting in 475 raw features, which are the input to a 4-layer MLP with the bottleneck architecture [13]. The size of the bottleneck layer is the desired number of features (39). A 3-dimensional pitch feature vector (pitch, Δ and $\Delta\Delta$ pitch) is combined with the other features, resulting in a total of 81 parameters (mlpplpf0).

MLP networks were trained for each of the three languages. The English network was was trained on over 2000 raw hours of conversational telephone speech for US English. The French MLP was trained about 1100 hours of French conversational data, most of the audio being associated with quick transcriptions. The Spanish MLP was trained on significantly less data (138 hours). All networks were trained using the scheme proposed in [14], reserving a portion of the data for cross-validation to monitor performance. Table 1 summarizes the characteristics of the MLP training. The MLP targets correspond to the individual phone states. It can be seen that the highest cross-validation accuracy is obtained for Spanish, and the lowest for French.

LANGUAGE	HOURS	#MLP	CV
	(RAW DATA)	TARGETS	ACCURACY
ENGLISH	2150	138	48.61%
FRENCH	1100	99	46.98%
SPANISH	138	90	53.93%

Table 1: MLP cross-validation frame accuracies.

Three PPRLM systems are compared. The first one uses CI phone models trained with PLP features (CI PLP) where as the second uses CD phone models (CD PLP) instead. More details about these systems can be found in [1] with minor modifications, the English recognizer has 38 phones (instead of 48) and all CD phone models uses 2000 contexts (instead of 3000). The third system uses CI phone models trained on the concatenated MLPPLPF0 features. For the three PPRLM systems, all phone recognizers for a specific language are trained on the same data and have the set of phonemes.

3. Experimental Setup and Results

3.1. Experimental set-up

The three PPRLM systems are evaluated on a closed-set language detection task, using the NIST LRE 2005 evaluation data sets[3]. There are 7 target languages and most of the eval segments are from the OHSU database. Target language phonotactic models are trained only with the CallFriend database. Table 2 specifies the amount of training data for each target language after removing non-speech segments. For each target language and phone recognizer pair, a 4-gram back-off phonotactic language model is estimated using Witten-Bell smoothing technique. Phone n-gram statistics are estimated from phone lattices. CMLLR adaptation is performed prior to phone lattice decoding. This adaptation is performed only for segments longer than 6 seconds.

LANGUAGE	HOURS (PROCESSED DATA)
ENGLISH	16.7
HINDUSTANI	9.2
JAPANEESE	8.6
KOREAN	7.9
MANDARIN	17.4
SPANISH	18.5
TAMIL	7.4

Table 2: The amount of training data (in hours) for each target language after removing automatically detected non-speech segments.

During test, language phonotactic scores are first combined and calibrated using an adapted Gaussian back-end followed by logistic regression [15]. The parameters of the fusion module are trained on a development set composed of four NIST data sets: LRE96 (dev& eval). LRE03 eval and LRE05 dev. No external data sources are used. The decision is made on the log likelihood ratio and performance is reported in terms of the $C_{avg}[\%]$ measure.

3.2. Using MLP features

The three systems were first evaluated on the $30s$ condition of the NIST LRE05 eval data. Table 3 reports the experimental results obtained by each individual PRLM component and by the combined PPRLM systems.

[3] http://www.nist.gov/speech/tests/lang/2005/

SYSTEM	FR	SP	EN	FUSION
CI PLP	8.0	6.4	7.5	4.7
CI PLP+CMLLR	6.5	5.4	5.5	3.7
CD PLP	5.0	3.9	4.8	3.0
+CMLLR	3.8	3.3	4.2	**2.5**
CI MLPPLPF0	5.4	5.0	5.3	2.8
+CMLLR	5.3	4.8	4.6	**2.4**
+CMLLR(MLP)	5.3	4.9	5.6	3.0
+CMLLR(PLPF0)	5.2	4.7	4.9	2.5

Table 3: Performance of different PPRLM systems in terms of $C_{avg}[\%]$ on the $30s$ condition of the NIST LRE 2005. Individual French (FR), Spanish (SP) and English (EN) PRLM components and result of fusion.

The first two rows report performances of CI PLP systems without (first row) and with (second row) CMLLR adaptation. CMLLR adaptation gives about 21% relative improvements. Compared to the adapted CI PLP system, using the CI MLPPLPF0 system without CMLLR adaptation (row 4), results in relative improvements of the individual PRLMs of 17%, 7.4% and 4%, for FR, SP, and EN respectively. Since the MLP is trained with input vectors of approximately $500ms$, this improvement can be largely attributed to the temporal information (context) that is better captured with the MLP features. Capturing such information is essential for language recognition task, as demonstrated with SDC features with acoustic approaches [6]. These improvements are less than those obtained with CD PLP system without CMLLR adaptation (row 2, Table 3). Using the latter models, the relative improvements of the FR, SP and EN PRLMs are, 23%, 28% and, 13%, respectively.

Interestingly, the results of the PPRLM systems show different behaviors. Compared to the CI PLP PPRLM system, the CI MLPPLPF0 system obtained a 24% relative improvement, which is larger than that obtained with the CD PLP system (19%). Thus it can be seen that the combination is more beneficial with the MLPPLPF0-based PPRLM system. In the PLP systems, all phone recognizers are using the same set of feature vectors (PLP) for phone lattice decoding.

In the MLPPLPF0 system, MLP features are generated using language specific networks, thus only half of the concatenated features are in common. It appears that the different MLP features provide some complementary information in the phone lattices from which phonotactics statistics are estimated. The lattices generated with the PLP-based systems appear to contain less complementary information. This leads to a larger improvement during phonotactic score combination. In other words, with the PLP PPRLM system, only phone recognizers are different, while in the MLPPPF0 PPRLM system, both the phone recognizers and features are different.

The CI MLPPLPF0 PPRLM system even performs better (7% relative) than the CD PLP PPRLM system with an average processing speed of about 6 times faster.

3.3. CMLLR adaptation

CMLLR adaptation is applied to reduce the mismatch between the adaptation data (the language training data) and the acoustic models (the acoustic phone models). This is performed by estimating a set of transformation matrices to maximize the likelihood of the adaptation data given the acoustic models. In this work, only one transformation is used. The statistics are collected from the one best hypotheses generated in a single decoding pass. In both CI and CD PLP systems, a full transformation matrix (39×40 parameters) is used. With the MLPPLPF0 system, and since the feature vector has 81 dimensions, estimating a full CMLLR matrix was found to be computationally prohibitive. (The time needed to estimate the CMLLR matrices was much higher than the decoding time, making the system quite inefficient). Instead, a block-diagonal transform is used. In this scheme, two sub-transformation matrices are estimated separately, one for the MLP features and the second for the PLPF0 features. The transforms are estimated using acoustic phone models trained separately on only MLP or PLPF0 features.

Using CMLLR adaptation, the relative improvements of the CD PLP and the CI MLPPLPF0 PPRLM systems (rows 3 and 5 in Table 3), over their corresponding unadapted systems are 17% and 14%, respectively.

In the MLP training, acoustic classes are discriminatively trained by adjusting the decision boundaries and mapping similar acoustic frames to the same region (class). This mapping is done independently of the nuisance factors (such as speaker variations and channel conditions) that are not relevant. In doing such a mapping, the MLP tries to normalize (remove) such information and exploiting the lexical context of the speech. The amount of the normalized information depends on how fine the choice of the acoustic targets. For example, when the targets represent phonemes, the amount of the information to be normalized is more important than with broad phone classes (such as fricatives, vowels, etc). Therefore the MLP actually does some of the work that the CMLLR adaptation tries to do. This explain why the CMLLR adaptation was more beneficial to the PLP system than to the MLPPLPF0 system, in particular for the individual PRLMs.

To support this idea, rows 6 and 7 of Table 3 report results with the MLPPLPF0 system when only one of the MLP or PLPF0 CMLLR matrices is used. Compared to the MLPPLPF0 system without adaptation, adapting only the MLP features degrades system performance, while adapting PLPF0 features improves system performance significantly. Interestingly, the latter system performs comparably with the system using the block-diagonal CMLLR transforms.

3.4. Results on short segments

As mentioned in the introduction, phonotactic approaches do not perform as well on short segments as they do for long segments. Some experiments were thus carried out to quantify the performance of the MLPPLPF0 system on shorter segments. The three PPRLM systems are compared on the 10 and 3 second conditions of the NIST LRE'05 data. CMLLR adaptation was applied for all systems. The results are reported in Table 4.

SYSTEM	10S	3S
CI PLP+CMLLR	8.7	18.2
CD PLP+CMLLR	6.2	14.6
CI MLPPLPF0+CMLLR	6.9	15.5

Table 4: $C_{avg}[\%]$ performances of different systems on the 10 and 3 second conditions of the NIST LRE 2005.

It can be seen that the MLPPLPF0 PPRLM system still outperforms the CI PLP (20% and 15% on the $10s$ and $3s$ conditions, respectively) but performs less well than the CD PLP system. It should be noted that the decoding parameters of the proposed system have not yet optimized, so better performance can be expected.

4. Conclusion

This paper has proposed the use of MLP features in combination with PLP features to build efficient phonotactic language recognition systems. The incorporation of MLP features allowed the development of a PPRLM system that has comparable performance to a state-of-the-art phonotactic system while running 6 times faster. These promising results encourage us to further explore MLP features for the language recognition task. Ongoing work aims at optimizing the system and evaluating it on other NIST LRE eval data sets. Initial results on the NIST LRE07 show that the proposed system outperforms the CMLLR adapted CD PLP system by 20% relative. Future work will also address finding an efficient way to incorporate MLP features within PPRLM system that uses context-dependent phone models to gain further improvements. Straightforward decoding with CD phone models and the concatenated features may improve system performance but at a high computational cost, incompatible with real-world applications.

REFERENCES

[1] 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, Brisbane, 2008.

[2] W. Shen and D. Reynolds. "Improved Phonotactic language recognition with acoustic adaptation" *InterSpeech'07*, pp. 358-361, Antwerp, 2007.

[3] J. L. Gauvain, A. Messaoudi and H. Schwenk, "Language Recognition Using Phone Lattices", *ICSLP'04,* Jeju, 2004.

[4] M.F. BenZeghiba, J.L. Gauvain and L. Lamel, "Fusing Language Information from Diverse Data Sources for Phonotactic Language Recognition" *To appear in Odyssey'12*, Singapore, 2012.

[5] A. Stolcke, M. Akbacak, L. Ferrer, S. Kajarekar, C. Richey, N. Scheffer, and E. Shriberg, "Improving Language Recognition with Multilingual Phone Recognition and Speaker Adaptation Transforms" *Odyssey'10*, pp. 256-262, Brno, 2010.

[6] B . Bielefeld "Language Identification using Shifted Delta Cepstrum" $14th$ *Annual Speech Research Symposium*, 1994.

[7] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohlar, R. J. Green D. Reynolds and J. R. Deller "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features" *ICSLP'02* pp. 89-92, 2002.

[8] P. Matejka, P. Shwarz, J. Cernocky and P. Chytil "Phonotactic Language Identification using High Quality Phoneme Recognition", *Interspeech'05* pp. 2237-2240, Lisbon, 2005.

[9] L. Lamel, J.-L. Gauvain, V.-B. Le, I. Oparin, and S. Meng. "Improved Models for mandarin Speech-to-Text Transcription." *ICASSP'11*, pp. 4660-4663, Prague, 2011.

[10] H. Hermansky, "Perceptual Linear Prediction (PLP) Analysis for Speech," *JASA*, **87**:1738-1752, April, 1990.

[11] P. Schwarz, P. Matějka, J. Černocky, "Towards Lower Error Rates In Phoneme Recognition," *TSD'04*, 465-472, Brno, 2004.

[12] J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, **37**:89-108, 2002.

[13] F. Grézl, P. Fousek, Optimizing Bottle-Neck Features for LVCSR, *ICASSP'08*, 4729-4732, Las Vegas, 2008.

[14] P. Fousek, L. Lamel, J.L. Gauvain, "On the Use of MLP Features for Broadcast News Transcription," *TSD08.* LNCS 5246/2008, 303.10, Springer Verlag, 2008.

[15] M.F. BenZeghiba, J.L. Gauvain and L. Lamel, "Language Score Calibration using Adapted Gaussian Back-end" *Interspeech'09*, pp. 2191-2194, Brighton, 2009.