Using Time-Synchronous Phone Co-occurrences in a SVM-Phonotactic Dialect Recognition System

Amparo Varona, Mikel Penagarikano, Luis Javier Rodriguez-Fuentes, German Bordel, Mireia Diez

GTTS, Department of Electricity and Electronics, ZTF/FCT
University of the Basque Country, UPV/EHU, Barrio Sarriena, 48940 Leioa, Spain
amparo.varona@ehu.es

Abstract

This paper presents a simple approach to phonotactic dialect recognition which uses lattices of time-synchronous phone co-occurrences at the frame level. In previous works, we successfully applied cross-decoder phone co-occurrences to improve performance in language recognition experiments on the 2007 NIST LRE database. We call phone co-occurrence to the simultaneous (time-synchronous) presence of two phone units coming from two different phone decoders. In this work, the approach is ported to a Dialect Recognition task based on the assumption that co-occurrences can better represent the tiny differences among the dialects. Besides, a slightly different approach is presented, based on the simultaneous presence of two phone units in the lattice produced by a single decoder (intra-decoder phone co-occurrences). For evaluating the approach, a choice of open software (Brno University of Technology phone decoders, HTK, SRILM, LIBLINEAR and *FoCal*) was used, and experiments were carried out on the Arabic dialects of the NIST 2011 LRE database. The approach based on cross-decoder phone co-occurrences outperformed the baseline phonotactic system, yielding around 8% relative improvement. The fusion of both systems yielded 7.31% EER and $C_{LLR} = 0.497$, meaning 19% relative improvement.

Index Terms: Phonotactic Dialect Recognition, Phone Co-occurrences, Phone Lattices, Support Vector Machines.

1. Introduction

In the last years, Spoken Language Recognition (SLR) has experienced great progress since it is demanded by many applications such as: spoken language translation, multilingual speech recognition, spoken document retrieval, etc. NIST Language Recognition Evaluations (LRE) [1] have significantly contributed to the development of SLR technology since 1996 until today. Dialect Recognition (DR) is considered an even more challenging problem, since differences among dialects are more subtle than those found among different languages. In the last NIST LREs, some DR tasks have been also included: Chinese in NIST 2007 LRE, English (American and Indian) in NIST 2007, 2009 and 2011 LRE, and Arabic (Iraqi, Levantine, MSA and Maghrebi) in the last NIST 2011 LRE.

The techniques applied to Dialect Recognition are usually ported from Language Recognition. Two main complementary approaches are typically used: *low level* acoustic modeling and *high level* phonotactic modeling. To model the target language, *low level* acoustic systems take information from the spectral characteristics of the audio signal as in [2], whereas *high level* phonotactic systems use sequences of phones produced by Parallel Phone Recognizers (PPR) as in [3, 4, 5]. An hybrid approach has been recently proposed in [6].

Nowadays, the most common phonotactic approach uses counts of phone n-grams to build a feature vector which feeds a classifier based on Support Vector Machines (SVM) [7]. System performance can be improved with the use of phone lattices instead of 1-best phone strings [8], since lattices provide richer and more robust information.

In previous works, we have presented a new approach to phonotactic language recognition which uses statistics of Cross-Decoder Phone Co-occurrences (CDPC) at the frame level starting from 1-best phone strings in [9], and from lattices in [10]. CDPC take into account the simultaneous (time-synchronous) presence of two phone units (co-occurences) coming from two different phone decoders. Not all the co-occurrences must be considered: in [10] we showed that using the 200 most likely co-occurences was a good compromise to get optimum performance. Experiments were carried out on the NIST 2007 LRE database yielded 15% of relative improvements.

In this work, this approach is ported to a Dialect Recogniton task. We start from the assumption that co-occurrences can better represent the tiny differences among dialects. Besides, a variant of the CDPC approach that can be applied using a single decoder, is presented: Intra-Decoder Phone Co-occurrences (IDPC) which take into account the simultaneous presence of two phone units in the phone lattice produced by a single decoder. Systems have been developed by means of open software (BUT phone decoders, HTK, SRILM, LIBLINEAR and *FoCal*) and evaluated on a relevant database: the Arabic dialects of the NIST 2011 LRE (Iraqi, Levantine, MSA and Maghrebi).

The rest of the paper is organized as follows. Section 2 presents the main features of the lattice-based phonotactic recognition system used as baseline in this work. Section 3 describes the proposed approach, based on the use of lattices to compute statistics of time-synchronous phone co-occurrences. The experimental setup is briefly described in Section 4. Results obtained in dialect recognition experiments on Arabic dialects of the NIST 2011 LRE are presented in Section 5. Finally, conclusions are outlined in Section 6.

2. Baseline: Phonotactic Recognition

A phonotactic dialect recognizer based on phone lattices and SVM scoring is used as baseline system. An energy-based voice activity detector is applied in first place, which splits and removes long-duration non-speech segments from the signals. Then, the Temporal Patterns Neural Network (TRAPs/NN) phone decoders, developed by the Brno University of Technology (BUT) for Czech (CZ), Hungarian (HU) and Russian (RU)

[11], are applied to compute phone lattices. Regarding channel compensation, noise reduction, etc. all the systems presented in this paper rely on the acoustic front-end provided by BUT decoders.

BUT decoders do not generate phone lattices but phoneme posterior probabilities, which are stored and later processed with the HVite decoder of HTK [12]. BUT decoders take into account three non-phonetic units: *int* (intermittent noise), *pau* (short pause) and *spk* (non-speech speaker noise) along with 42, 58 and 49 phonetic units of Czech, Hungarian and Russian respectively. For each unit, a three-state model is used, so three posterior probabilities per frame are calculated and stored.

Before generating phone lattices, non-phonetic units *int, pau* and *spk* are integrated into a single 9-state model (*pau*). After that, the number of units is 43 for Czech, 59 for Hungarian and 50 for Russian. Then, posterior probabilities are used as input to HVite to produce phone lattices, which encode multiple hypotheses with acoustic likelihoods. Finally, the *lattice-tool* of *SRILM* [13] is used to produce the expected counts of phone n-grams.

3. Time-Synchronous Phone Co-occurrences

Two different approaches are considered: (1) Cross-Decoder Phone Co-occurrences (CDPC) [10] which take into account the simultaneous (time-synchronous) presence of two phone units (co-occurences) coming from two different phone decoders; and (2) Intra-decoder Phone Co-occurrences (LDPC) which take into account the simultaneous presence of two phone units in the phone lattice of a single decoder.

3.1. Cross-Decoder Phone Co-occurrences

Let us consider a choice of two decoders A and B from the set of 3 possible decoders (CZ, HU and RU). Lattices of time-synchronous cross-decoder phone co-occurrences can be obtained by composing the posterior probabilities of two phone models i, j (i from decoder A and j from decoder B) at each frame t, thus creating a single unit on the lattice which represents the time-synchronous co-occurrence of both phones.

BUT decoders produce a sequence of numbers representing the posterior probabilities $p_{i,s}^t$ for each one of the three states s of each phone i at each frame t, encoded in the following way:

$$x(p_{i,s}^t) = \sqrt{-2 \log p_{i,s}^t} \tag{1}$$

To combine posterior probabilities of two phones, i from decoder A and j from decoder B, at each state s and each frame t, the following expression can be applied:

$$\begin{aligned} x(p_{ij,s}^t) = x(p_{i,s}^t, p_{j,s}^t) &= \sqrt{-2 \log (p_{i,s}^t p_{j,s}^t)} \\ &= \sqrt{-2(\log p_{i,s}^t + \log p_{j,s}^t)} \\ &= \sqrt{x^2(p_{i,s}^t) + x^2(p_{j,s}^t)} \end{aligned} \tag{2}$$

Therefore, the new composite model ij, corresponding to the time-synchronous co-occurrence of phones i and j, has the same number of states than the models i and j. All the phonetic units of decoder A can be combined with all the phonetic units of decoder B, resulting composite models of three states. However, the *pau* model, which is always present in both decoders for any choice of A and B, is not taken into account.

As it was shown in a previous work [10], not all the phone co-occurrences must be considered, but only the most likely. To determine which 2-phone combinations should be used, the sum of posterior probabilities for each composite unit ij was calculated on the entire training set of the Arabic NIST 2011 LRE database (see Subsection 4.1 for details) in the following way:

$$p_{ij} = \sum_{s=1}^{S} \sum_{t=1}^{T} p_{ij,s}^t \tag{3}$$

where S is the number of states corresponding to each composite unit ij and T is the total number of frames in the training database. Once p_{ij} has been calculated for all the combinations, a ranked list is created by sorting the values of p_{ij} from highest to lowest. In this work, a set of 200 co-occurrences has been considered a good compromise to get optimum performance. We have found that this set of 200 co-occurrences differs by more than 25% from the set of 200 units that was chosen by using general SLR databases, that is, it seems that the set of co-occurrences has adapted to the language under study (Arabic).

Once the new sequence of posterior probabilities representing the phone co-occurrences is obtained, the HVite decoder can be used to get the composite lattice. The symbols associated to the arcs of the lattice represent the co-occurrence of two units, but the computational process is exactly the same as in the baseline system.

3.2. Intra-Decoder Phone Co-occurrences

Let us consider one decoder A from the set of 3 possible decoders (CZ, HU and RU). Lattices of time-synchronous intra-decoder phone co-occurrences can be obtained by composing the posterior probabilities of two phone models i, j (both i and j from decoder A) at each frame t, thus creating a single unit on the lattice which represents the time-synchronous co-occurrence of both phones. The process is exactly the same as that used to generate cross-decoder phone co-occurrences (see Subsection 3.1).

4. Experimental Setup

4.1. Training, development and test datasets

In the NIST 2011 LRE, 24 target languages were considered, and among them, four Arabic dialects: Iraqi, Levantine, MSA and Maghrebi. Development data specifically collected for the NIST 2011 LRE was sent to participants [14], including 100 30-second segments for each of the four Arabic dialects. For a better coverage, we split Arabic subsets into two disjoint subsets (each having approximately half the segments for each dialect): one half was used to train specific models for the Arabic dialects, and the other was used to estimate backend and fusion parameters.

To train more robust models, we added data from databases distributed by the LDC, some of them containing conversational telephone speech (LDC2006S45 for Arabic Iraqi and LDC2006S29 for Arabic Levantine). We also added data extracted from wideband broadcast news recordings, dowsampling them to 8 kHz and applying the Filtering and Noise-adding Tool[1] FANT to simulate a telephone channel. Arabic MSA was extracted from Al Jazeera broadcasts included in the Kalaka-2 database created for the Albayzin 2010 LRE [15]. Finally, broadcasts were also captured from video archives in TV

[1] http://dnt.kr.hsnr.de/download.html

websites to get speech segments in Arabic Maghrebi (Arrabia TV, http://www.arrabia.ma). TV broadcasts were fully audited, so that only reasonably clean speech segments were selected for training. Evaluation was carried out on the Arabic signals of the 2011 LRE evaluation corpus, specifically on the 30-second, closed-set condition. Table 1 summarizes the datasets used in the experiments.

Table 1: Dialect data used in the experiments.

Arabic Dialects	Hours		# 30s cuts	
	Train		Devel	Eval
	Other Sources	NIST LRE 2011		
Iraqi	20.24	0.48	48	308
Levantine	27.56	0.47	49	308
MSA	1.87	0.47	51	306
Maghrebi	1.79	0.41	54	305

4.2. Evaluation measures

In this work, systems will be compared in terms of: (1) Equal Error Rate (EER); and (2) the so called C_{LLR} [16], an alternative performance measure used in NIST evaluations. We internally consider C_{LLR} as the most relevant performance indicator, because it allows to evaluate system performance globally by means of an application independent single numerical value. C_{LLR} does not depend on application costs; instead, it depends on the calibration of scores, an important feature of detection systems. C_{LLR} has higher statistical significance than EER, since it is computed starting from verification scores (in contrast to EER, which depends only on Accept/Reject decisions).

4.3. SVM modeling

All systems developed in this work follow a SVM phonotactic approach. SVM vectors consist of counts of features representing the phonotactics of an input utterance: phone 3-grams (baseline), 3-grams of cross-decoder phone co-occurrences (co-oc X-Y) and 3-grams of intra-decoder phone co-occurrences (co-oc X-X). A sparse representation was used, which involved only the most frequent features. That is, instead of using a full space representation, features were ranked according to their counts on the training dataset using a feature selection algorithm based on frequency [17], and only those with the $M = 20000$ highest counts were considered. Counts were stacked in a single vector and an L2-regularized L1-loss Support Vector Machine (SVM) classifier was estimated and applied, by means of LIBLINEAR [18], which was modified by adding some lines of code to compute regression values. Finally, systems were built by fusing the scores of three calibrated SVM-based phonotactic subsystems (one per phone decoder). The *FoCal* toolkit was used for calibration and fusion (see [19] for details).

5. Experimental Results

Phonotactic systems described in Sections 2 and 3 were developed and evaluated on the Arabic Dialects of the NIST 2011 LRE (see Subsection 4.1 for details). Table 2 shows EER and C_{LLR} performance in dialect recognition experiments applying the baseline system, the proposed cross-decoder phone co-occurrence systems (co-oc X-Y) and intra-decoder phone co-occurrence systems (co-oc X-X). Note that we call *system* to the fusion of three subsystems, each corresponding to one phone decoder in the baseline and co-oc X-X systems (CZ, HU, RU and CZ-CZ, HU-HU, RU-RU respectively) and to a 2-decoder choice in the co-oc X-Y systems (CZ-HU, CZ-RU, HU-RU).

Baseline performance (9.03% EER and 0.613 C_{LLR}) was improved by the use of the proposed co-oc X-Y system (8.29% EER and 0.569 C_{LLR}) meaning around 8% relative improvement. When the proposed co-oc X-X was applied, a slight degradation was observed with regard to the baseline system (9.69% EER and 0.646 C_{LLR}). It seems that cross-decoder phone co-occurrences provide more information than the phone co-occurrences computed on a single decoder. This result was somehow expected, since time-synchronous phone likelihoods in a single decoder could probably be strongly correlated, i.e. their distribution may depend more on the decoder than on the language/dialect.

Fusions involving the baseline and the two co-occurrence systems are also presented in Table 2. Fusions led to better performance in all cases. Best performance was achieved by fusing the baseline and co-oc X-Y systems (7.31% EER and $C_{LLR} = 0.497$, meaning 19% relative improvement).

Table 2: Performance (EER and C_{LLR}) for the baseline system, the proposed cross-decoder phone co-occurrence (co-oc X-Y) and intra-decoder phone co-occurrence (co-oc X-X) systems, and different fusions.

System	EER	C_{LLR}
(1) CZ	13.01	0.881
(2) HU	14.24	0.920
(3) RU	13.51	0.878
Baseline = (1)+(2)+(3)	9.03	0.613
(4) CZ-HU	11.94	0.805
(5) CZ-RU	12.20	0.805
(6) HU-RU	10.75	0.733
Co-oc X-Y = (4)+(5)+(6)	8.29	0.569
(7) CZ-CZ	14.99	0.969
(8) HU-HU	14.81	0.974
(9) RU-RU	14.39	0.909
Co-oc X-X = (7)+(8)+(9)	9.69	0.646
Baseline + Co-oc X-X	8.60	0.557
Baseline + Co-cc X-Y	7.31	0.497
Co-oc X-X + Co-oc X-Y	7.64	0.527
Baseline + Co-oc X-X + Co-oc X-Y	7.39	0.492

For the sake of completeness, Table 3 and Table 4 show the performance of subsystems and fusions if only phone lattices of one-decoder or two-decoders were available. First, if only one decoder was available, the obtained results would be those of Table 3. Co-oc X-X subsystems led to worse performance than baseline subsystems, which is fully in line with previous results. But the fusion of baseline and co-oc X-X subsystems led to better performance: 4% and 9% for CZ, 8.5% and 8% for HU and 7.5% and 9.5% for RU relative improvements in EER and C_{LLR}, respectively.

Table 3: Performance (EER and C_{LLR}) for the baseline subsystems (CZ, HU, RU), the proposed intra-decoder phone co-occurrence subsystems (co-oc X-X: CZ-CZ, HU-HU, RU-RU) and fusions of baseline systems and co-oc X-X subsystems.

System	EER	C_{LLR}
CZ	13.01	0.881
CZ-CZ	14.99	0.969
CZ + CZ-CZ	12.52	0.801
HU	14.24	0.920
HU-HU	14.81	0.974
HU + HU-HU	13.02	0.842
RU	13.51	0.878
RU-RU	14.39	0.909
RU + RU-RU	12.53	0.794

Then, if only two decoders were available, the obtained results would be those of Table 4. Again, co-oc X-Y subsystems led to worse performance than baseline systems (fusion of two baseline subsystems), which is also consistent with previous results. But the fusion of the baseline and the co-oc X-Y subsystems led to better performance: 14% and 13% for (CZ, HU), 11% and 11% for (CZ, RU) and 20% and 16.5% for (HU,RU) relative improvements in EER and C_{LLR}, respectively.

Table 4: Performance (EER and C_{LLR}) for the baseline subsystems (CZ+HU, CZ+RU, HU+RU), the proposed cross-decoder phone co-occurrence subsystems (co-oc X-Y: CZ-HU, CZ-RU, HU-RU) and fusions of baseline and co-oc X-Y subsystems

System	EER	C_{LLR}
CZ + HU	10.65	0.699
CZ-HU	11.94	0.805
CZ + HU + CZ-HU	9.11	0.609
CZ + RU	10.26	0.696
CZ-RU	12.20	0.805
CZ + RU + CZ-RU	9.12	0.618
HU + RU	10.56	0.699
HU-RU	10.75	0.733
HU + RU + HU-RU	8.37	0.584

We have carried out another series of experiments on the American vs. Indian English dialects. Training and development data were limited to those distributed by NIST to 2007 LRE participants: 10 conversations per language were randomly selected for development (204 American English, 84 Indian English and 174 Hindi 30-second cuts) and the remaining conversations were used for training (130.7 hours of American English, 13 hours of Indian English and 59.3 hours of Hindi). We evaluated these systems on the official 2007 NIST LRE Test Set (30-second task) including 80 American English and 160 Indian English signals. The baseline system yielded 7.81% EER, the co-oc X-Y system yielded 7.34% EER and the fusion of both yielded 7.03% EER meaning a 10% relative improvement and better performance (on the same task) than that reported by state-of-the-art phonotactic systems [3], acoustic systems [2] and hybrid systems [6], thus supporting the use of cross-decoder dependencies for dialect recognition. However, due to the small number of test trials, these results are not as significant as those presented above for the Arabic dialects.

6. Conclusions

In this paper, the latest developments under two approaches using lattices of cross-decoder and intra-decoder time-synchronous phone co-occurrences in SVM-based phonotactic dialect recognition have been presented and evaluated. The proposed approaches rely on the assumption that co-occurrence information is somehow specific to each dialect and represents just a means to extract more information from existing decodings. The proposed cross-decoder co-occurrence approach outperformed the baseline phonotactic system yielding around 19% relative improvement in terms of C_{LLR}. The fusion of the baseline system and the cross-decoder co-occurrence approach yielded 7.31% EER and $C_{LLR} = 0.497$, meaning also 19% relative improvement. This support our assumption that the use of co-occurrences help to better represent the tiny differences among dialects.

7. Acknowledgments

This work has been supported by the University of the Basque Country, under grant GIU10/18 and project US11/06 and the Spanish MICINN, under *Plan Nacional de I+D+i* (project TIN2009-07446, partially financed by FEDER funds). Mireia Diez is supported by the Department of Education, Universities and Research of the Government of the Basque Country

8. References

[1] *NIST LRE*, http://www.itl.nist.gov/iad/mig/tests/lre/.

[2] P. A. Torres-Carrasquillo, D. E. Sturim, D. A. Reynolds, and A. McCree, "Eigen-channel compensation and discriminatively trained gaussian mixture models for dialect and accent recognition," in *INTERSPEECH*, 2008, pp. 723–726.

[3] F. S. Richardson, W. Campbell, and P. A. Torres-Carrasquillo, "Discriminative n-gram selection for dialect recognition," in *INTERSPEECH*, 2008, pp. 192–195.

[4] M. Akbacak, D. Vergyri, A. Stolcke, N. Scheffer, and A. Mandal, "Effective arabic dialect classification using diverse phonotactic models," in *INTERSPEECH*, 2011, pp. 737–740.

[5] N. F. Chen, W. Shen, J. P. Campbell, and P. A. Torres-Carrasquillo, "Informative dialect recognition using context-dependent pronunciation modeling," in *ICASSP*, 2011.

[6] F. Biadsy, J. Hirschberg, and D. P. W. Ellis, "Dialect and accent recognition using phonetic-segmentation supervectors," in *INTERSPEECH*, 2011, pp. 745–748.

[7] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, no. 2–3, pp. 210–229, 2006.

[8] W. M. Campbell, F. Richardson, and D. A. Reynolds, "Language recognition with word lattices and support vector machines," in *ICASSP*, Honolulu, 2007, pp. 15–20.

[9] M. Penagarikano, A. Varona, L. J. Rodriguez-Fuentes, and G. Bordel, "Improved modeling of cross-decoder phone co-occurrences in svm-based phonotactic language recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 8, pp. 2348–2363, 2011.

[10] A. Varona, M. Peñagarikano, L. J. Rodríguez, and G. Bordel, "On the use of lattices of time-synchronous cross-decoder phone co-occurrences in a svm-phonotactic language recognition system," in *INTERSPEECH*, 2011, pp. 2901–2904.

[11] P. Schwarz, *Phoneme recognition based on long temporal context*, 2008, ph. D. Fac. of Information Technology BUT.

[12] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Lui, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (Version 3.4)*, Cambridge, 2006.

[13] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proceedings of ICSLP*, November 2002, pp. 257–286.

[14] M. Penagarikano, A. Varona, L. Rodriguez-Fuentes, M. Dez, and G. Bordel, "University of the basque country (ehu) systems for the 2011 nist language recognition evaluation," in *Proceedings of the NIST 2011 LRE Workshop*, Atlanta (USA), 6-7 december 2011.

[15] L. Rodriguez-Fuentes, M. Penagarikano, A. Varona, M. Dez, and G. Bordel, "Kalaka-2: A tv broadcast speech database for the recognition of iberian languages in clean and noisy environments," in *LREC*, Istambul (Turkey), 21-27 May 2012.

[16] N. Brümmer and J. A. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2–3, pp. 230–275, 2006.

[17] M. Penagarikano, A. Varona, L. Rodríguez-Fuentes, and G. Bordel, "A dynamic approach to the selection of high-order n-grams in phonotactic language recognition," in *Proceedings of ICASSP*, Prage, Czech Republic, 2011.

[18] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[19] N. Brümmer and D. van Leeuwen, "On calibration of language recognition scores," in *Proceedings of Odyssey - The Speaker and Language Recognition Workshop*, 2006, pp. 1–8.