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

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

Arabic Dialect Identification – 'Is the Secret in the Silence?' and Other Observations

*Hynek Bořil, Abhijeet Sangwan, John H.L. Hansen**

[1]Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering,
University of Texas at Dallas, Richardson, Texas, U.S.A.
{hynek,abhijeet.sangwan,john.hansen}@utdallas.edu

Abstract

Conversational telephone speech (CTS) collections of Arabic dialects distributed trough the Linguistic Data Consortium (LDC) provide an invaluable resource for the development of robust speech systems including speaker and speech recognition, translation, spoken dialogue modeling, and information summarization. They are frequently relied on also in language (LID) and dialect identification (DID) evaluations. The first part of this study attempts to identify the source of the relatively high DID performance on LDC's Arabic CTS corpora seen in recent literature. It is found that recordings of each dialect exhibit unique channel and noise characteristics and that silence regions are sufficient for performing reasonably accurate DID. The second part focuses on phonotactic dialect modeling that utilizes phone recognizers and support vector machines (PRSVM). A simple *N*-gram normalization of PRSVM input supervectors utilizing hard limiting is introduced and shown to outperform the standard approach used in current LID and DID systems.
Index Terms: Arabic dialect identification, channel characteristics, LDC corpora, PRSVM

1. Introduction

With the increasing presence of speech enabled systems in the multicultural setting, language and dialect identification (LID and DID) plays a crucial role in directing speech input to corresponding language- or dialect-specific acoustic and linguistic models (phonology, lexicon, grammar). The fast progress in the field of LID and DID is well documented in outcomes from the periodic NIST LRE campaigns [1].

Current state-of-the-art DID systems adopt many approaches previously developed for speaker and language recognition. In particular, short-term cepstral features with shifted delta cepstra (SDC) and Gaussian mixture modeling (GMM) [2–4], phonotactic models utilizing parallel phone recognizers and language modeling (PPRLM) [5–8], and phone recognizers with *N*-grams modeled by SVM classifiers (PRSVM) [9] are frequently used in state-of-the-art schemes.

In a recent study [10], SDC–GMM, PPRLM, and PRSVM systems were successfully applied in the task of Arabic dialect identification. The evaluations were performed on four Arabic dialects captured in the conversational telephone speech (CTS) databases available through the Linguistic Data Consortium (LDC): Levantine Arabic CTS, Iraqi Arabic CTS, Gulf Arabic CTS, and Egyptian CALLHOME/CALLFRIEND. The study [10] notes that the Egyptian data collection preceded the other databases and as a result might exhibit different channel characteristics versus the rest of the databases. The authors argue that proceeding with evaluations on this dataset is still useful as it allows for performance comparisons with previous efforts such as those reported in [11, 12].

In the first part of this study, channel and noise characteristics of selected LDC Arabic dialect CTS corpora are analyzed and found to be unique and fairly distinctive for each dialect corpus. As a consequence, silence segments are found to carry sufficient information to perform a reasonably accurate DID. It is also demonstrated that performing channel normalization may, to a large extent, help equalize channel differences between the dialect databases, but this is not sufficient in addressing other, most likely noise-related non-speech 'dialect cues' present in the recordings. In the second part of our study, a normalization of PRSVM input supervectors is proposed and evaluated alongside with the standard normalization on the LDC corpora as well as on an in-house Pan-Arabic corpus.

*This project was funded by AFRL under contract FA8750-12-1-0188 (Approved for public release, distribution unlimited: 88ABW-2012-1810), and partially by the U.S. Army through a subcontract to Li Creative Technologies, Inc. under W15P7T-10-C-B611, and by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

2. Corpora

This study uses two sets of Arabic dialect data. The first is represented by LDC's Arabic conversational telephone speech corpora: Iraqi Arabic CTS (IRQ), Gulf Arabic CTS (GLF), Arabic CTS Levantine Fisher Training Data Set 3 (LEV), and CALLHOME and CALLFRIEND Egyptian Arabic Speech (EGY). It is noted that [10–12] used Levantine Arabic CTS (LDC2007S01) instead of the Fisher corpus. While this represents a difference between our study and the past literature, it is assumed that the observations made in the following sections do apply also to the mentioned studies as 3 out of 4 dialect sets are overlapping.

The second data set is drawn from the Pan-Arabic corpus [13] provided by our sponsor. The corpus consists of Arabic dialect data from five different regions, including United Arab Emirates (AE), Egypt (EGY), Iraq (IRQ), Palestine (PS), and Syria (SY). Each dialect set contains 100 speakers (genders balanced). In each session, two speakers complete four combined conversational recordings. A lapel microphone is used in conversational recording for each speaker per conversation.

All speech recordings were segmented using an energy-based voice activity detector (VAD) which uses adaptive thresholding derived from the dynamic range of the energy envelope. To prevent possible dropping of low-energy consonants, time boundaries of the speech islands found by VAD are expanded by 0.4 sec. in both directions. This ensures the presence of a small portion of silence at the beginning and end of every speech island and is expected to benefit accuracy of phone recognition in a phonotactic DID by preserving natural transitions between silence and speech. The speech islands are concatenated into approximately 11–12 seconds long speech *chunks*. The amount of training and evaluation data for both data sets is shown in Table 1. Training and test sets comprise different speaker sessions.

	LDC Corpora				In-House Pan-Arabic			
	GLF	IRQ	LEV	EGY	PS	IRQ	SY	EGY
Train Set (Hrs)	32.7	16.1	11.9	33.9	10.6	9.3	10.8	9.9
Test Set (Hrs)	2.0	2.3	1.6	10.1	2.8	2.7	2.5	2.6
Avg. Chunk Length	11.3 sec				11.9 sec			

Table 1: Data set content.

3. Analysis of LDC Corpora

In the preliminary experiments involving a naïve maximum likelihood GMM classifier, separate 32-mixture GMMs were trained for each of the 4 dialects captured in the LDC data set. The number of training chunks was as follows: Iraqi (5075), Gulf (10526), Levantine (3771),

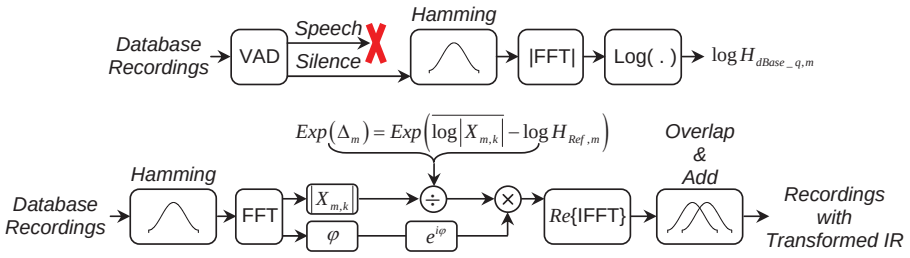

Figure 2: Block scheme of channel estimation, normalization, and signal reconstruction in time domain.

and Egyptian (10628). The signal was parameterized with an MFCC-like front-end where the triangular Mel frequency filterbank was replaced by 20 non-overlapping rectangular filters uniformly spread in linear frequency [14]. Thirteen static, delta, and delta-delta cepstral coefficients were extracted using 25 ms/10 ms windowing. The confusion matrix for an in-set dialect identification (pick 1-out-of-4) on the level of individual speech chunks is shown in Table 2. It can be seen that in spite of the simplicity of the system, the initial performance on short speech chunks is relatively high compared to the chance performance (25 %). In order to verify to what extent can the DID performance be attributed

	Assigned Dialect (Speech Chunks)				Acc (%) Avg 82.0
Ground Truth	Gulf	Iraqi	Levantine	Egyptian	
Gulf	**510**	120	4	1	80.3
Iraqi	184	**527**	1	2	73.8
Levantine	120	10	**370**	**0**	74.0
Egyptian	8	0	0	**3174**	99.7

Table 2: GMM-based DID on speech chunks.

to the linguistic content present in the recordings, another experiment, where pure silence chunks were used both for GMM training and evaluation, was conducted. As can be seen in Table 3, the average dialect classification accuracy increased from 82.0 % seen for speech chunks, to 83.3 % using silence chunks. This suggests that the silence regions themselves carry sufficient information for identifying the database origin and, in the case of the simplistic GMM classifier, presence of speech is actually not helpful to the task. Since the individual dialect databases

	Assigned Dialect (Silence Chunks)				Acc (%) Avg 83.3
Ground Truth	Gulf	Iraqi	Levantine	Egyptian	
Gulf	**260**	78	0	0	76.9
Iraqi	96	**228**	0	0	70.4
Levantine	24	1	**158**	1	85.9
Egyptian	0	0	0	**1973**	100

Table 3: GMM-based DID on silence chunks.

were acquired in separate efforts, it can be expected that channel characteristics captured in the recordings may vary and contribute to the identification of the data origin. It is noted that the concern expressed by [10] about the possible different acoustic characteristics captured in the Egyptian corpus compared to the other LDC corpora is confirmed by Table 3 as the Egyptian silence chunks are distinguished from the other data sets with the highest (100 %) accuracy. To further verify the hypothesis about the channel differences across the corpora, a channel estimation procedure was performed as depicted in the upper part of Fig. 2. For each database, the long-term channel transfer function was estimated by averaging short-term log-amplitude spectra of silence segments in the recordings. During the procedure, all segments containing either digital silence or having some of the energy spectrum bins equal to zero were omitted from the estimation. Average transfer function estimates for the four dialect recordings are shown in Fig. 1. The plots are accompanied by dashed lines representing intervals of $\pm 5\sigma$ ($\pm\sigma$ interval plots were below eye resolution). It can be seen that the chan-

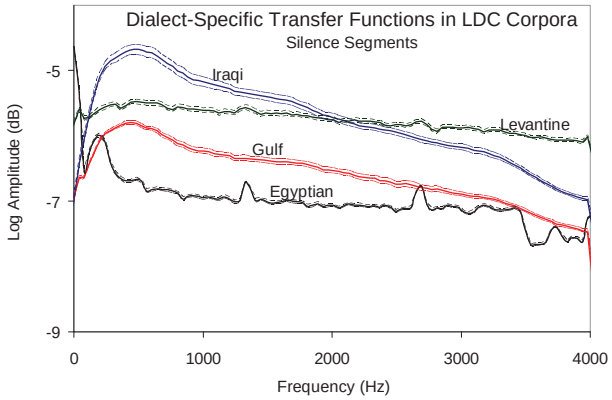

Figure 1: Arabic dialect-specific channel characteristics in LDC corpora estimated as long-term averages of log amplitude spectra in silence segments. Dashed lines – intervals of $\pm 5\sigma$.

nel characteristics are very consistent across each dialect recordings and fairly distinctive between dialects. In order to equalize database channel differences, a normalization procedure was implemented (see bottom part of Fig. 2). While the normalization could be conveniently applied directly on the cepstral coefficients, our goal is to reconstruct the normalized time-domain speech samples that could be later processed by any DID scheme of choice (e.g., utilizing embedded phone recognizers that perform their independent feature extraction). While the channel estimation was conducted in the log-spectral domain, the normalization is performed in the linear amplitude spectrum by dividing the actual short term amplitude spectrum by the delta-channel transfer function. This allows for normalization of all segments, including those in which some of the spectral bins are zero. The delta-channel transfer function Δ_m, where m is the index of spectral bin, is estimated as a difference between the average of non-zero short-term spectra of noise and target channel. The Iraqi transfer function was chosen as a target and all recordings were normalized towards it (including the Iraqi recordings, to ensure that all samples went through the same processing chain). Standard overlap-and-add technique utilizing a Hamming window shifted with an overlap step of 25 % is used in the time-domain signal reconstruction. The estimated average transfer functions of the normalized silence segments are shown in Fig. 3.

It can be seen that as a result of the normalization, the transfer functions migrate towards the target Iraqi channel transfer function. In an informal perceptual test, the reconstructed recordings did not contain any perceivable signal processing artifacts (such as loudness bursts, speech distortion, etc.). At the same time, the change in the channel characteristics was well audible, causing an impression the speech was acquired by different types of microphones or produced through different loudspeakers. Although the channel normalization seems to be successful,

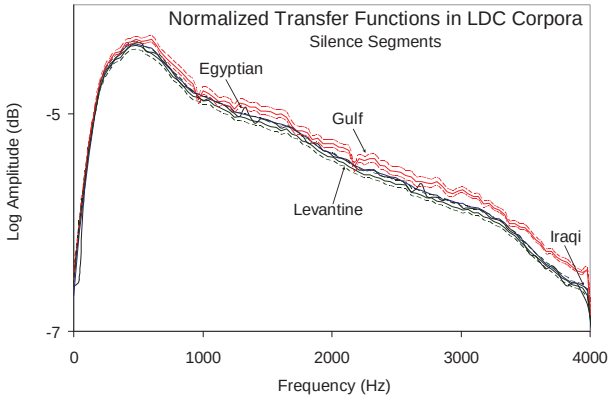

Figure 3: Normalized channel characteristics estimated from silence segments. Dashed lines – intervals of $\pm 5\sigma$ (depicted for clarity only for Gulf and Levantine).

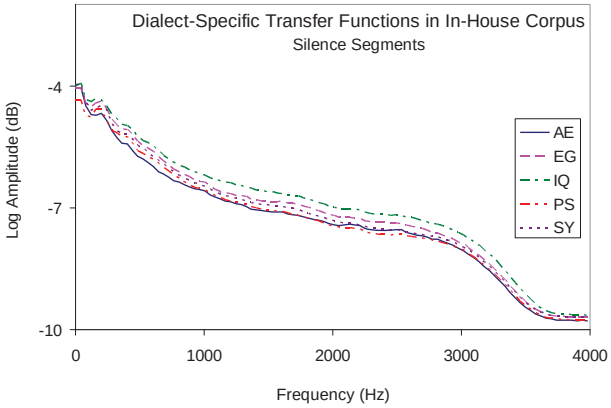

Figure 5: Arabic dialect-specific channel characteristics in in-house Pan-Arabic corpus estimated as long-term averages of log amplitude spectra in silence segments.

an experiment on the normalized silence chunks resulted in similar high DID accuracy as seen prior to normalization. In parallel, a repeated experiment on the unprocessed silence segments, where cepstral mean normalization was applied in the feature extraction front-end was also performed, providing comparable results to those on the normalized audio recordings. This suggests that the silence samples contain additional strong database-specific cues. Fig. 4 shows distributions of the first cepstral coefficient c_1 extracted from 100 silence chunks per dialect, prior and after the channel normalization. It can be seen that while the distribution means are, as a result of the normalization, aligned, the distribution contours are fundamentally different. Considering stationarity of the channel characteristics as suggested by Fig. 3, the variance and fine details in the distribution contours can be attributed to the noise present in the recordings. This suggests that additional signal processing will be necessary to equalize the non-linguistic content. It is noted that the CTS samples contain numerous non-stationary noises (random bursts of electrical noise, handset noises) and noise suppression techniques developed for stationary noises may not be as effective.

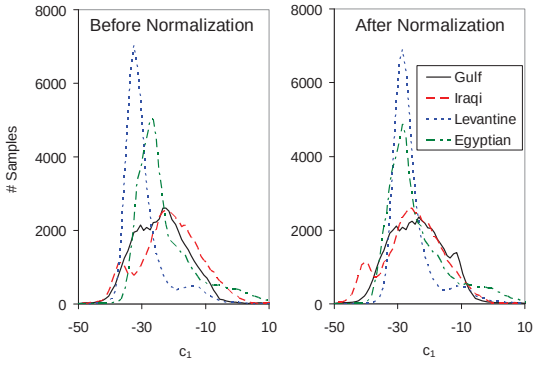

Figure 4: Distribution of c_1 before and after channel normalization; extracted from 100 silence chunks.

For a comparison, an identical GMM-based classification procedure was repeated also for silence segments from the in-house Pan-Arabic corpus. A classification accuracy in a four-way task on PS, SY, IRQ, and PS silence chunks (AE was omitted to mimic the complexity of the LDC task) yielded a slightly below chance (24.7 %) accuracy. This suggests that the acoustic characteristics of the silence segments here are much more consistent across the dialects. Figure 5 details corresponding long-term average transfer functions estimated from the silence segments (including AE). It can be seen that the transfer functions in this case exhibit much lower variance across dialects compared to the LDC corpora. This can be attributed to the fact that all dialect samples here were recorded in a single effort using a unified framework.

4. PRSVM System

In this section, a phonotactic dialect recognition system that combines phone recognition and *N*-gram modeling via SVM classifiers (PRSVM) is tested on the LDC and in-house Pan-Arabic dialect sets described in Sec. 2. The purpose is to evaluate the impact of the presence of channel/noise dialect cues on the PRSVM system. New normalization of the SVM input supervectors is also presented.

Phone Recognizer	AvgEER (%) (1-vs-3)							
	LDC Corpora				In-House Pan-Arabic			
	GBF +Log	GBF +Sig.	Sig.	Hard	GBF +Log	GBF +Sig.	Sig.	Hard
Czech	21.1	19.0	19.9	18.7	35.2	33.7	34.3	33.6
English	7.7	7.1	7.0	6.8	33.6	33.2	33.8	33.3
German	19.8	15.8	15.5	14.7	34.9	30.6	31.1	30.4
Hindi	16.7	14.2	14.5	13.8	31.6	29.4	30.3	29.4
Hungarian	16.6	14.9	15.0	14.8	36.6	35.0	35.4	34.9
Japanese	15.6	14.8	14.2	14.2	31.0	30.5	30.3	30.3
Mandarin	16.8	13.5	13.2	12.7	31.5	29.7	30.5	29.6
Russian	21.1	19.3	20.1	18.9	36.7	35.7	36.1	35.6
Spanish	19.0	17.3	16.7	16.1	33.9	33.1	34.7	33.4
AvgErr (%)	17.2	15.1	15.1	**14.5**	33.9	**32.3**	32.9	**32.3**

Table 4: Average EER (*AvgEER*) in PRSVM setup with various phone recognizers. Comparison of PRSVM input supervector normalizations (GBF – normalization of local bigram frequency (*LBF*) with a square root of global bigram frequency).

Our PRSVM implementation follows the ones in [4, 9]. A set of 9 phone recognizers developed by Brno University of Technology (BUT) [15] is evaluated in the PRSVM system: English, Czech, Hungarian, Russian, German, Hindi, Japanese, Mandarin, and Spanish. Each input speech sample is phone-decoded using a phone recognizer of choice. The recognizer output is parameterized by *N*-gram relative frequencies (number of occurrences normalized by the number of *N*-grams). *N*-gram relative frequencies are scaled by the inverse square root of the global frequency of the corresponding *N*-grams, following [4, 9], processed by a log squashing function $g_j = \log(x) + 1$ [16], and stacked into a supervector where each dimension represents a normalized frequency of a particular *N*-gram. The supervectors are in general sparse since the decoded sample is likely to contain only a few of all the possible *N*-grams [9]. In our implementation, the log squashing function, as well as its alternatives, are applied only to the non-zero dimensions of the supervectors. The supervectors extracted from the training set are used to train binary SVM classifiers, where one class represents a target

dialect and the anticlass the remainder of the dialects. This yields a set of 4 SVM classifiers.

Following [10], the experimental results are presented in the form of average EER of the four SVMs (the task is to decide whether the sample contains a claimed dialect or the remaining dialects – pick 1-vs-3). The results for PRSVMs utilizing various phone recognizers and bigram modeling are summarized in Table 4. In our preliminary experiments, bigram-based PRSVM outperformed a trigram setup, hence, bigram frequencies are used in the following experiments. The column 'GBF+Log', where GBF stands for global bigram frequency, represents a standard normalization as described above. In the adjacent column 'GBF+Sig', a sigmoid function is applied instead of the logarithm. Sigmoid represents a popular choice for activation functions in neural networks and provides, among other properties, an attractive means to compress *outlier* sample amplitudes into a compact dynamic range. The following column 'Sig' represents a case where GBF is excluded from the normalization and only the sigmoid is applied on relative bigram frequencies (denoted local bigram frequencies; LBF). We have observed that in the 'GBF+Sig'-based setups, the non-zero scores tend to be compressed to a numerically coherent cluster. This led to an idea to substitute all non-zero relative frequencies by a fixed constant – hard limiting (column 'Hard'). The constant was experimentally set to 0.3. Table 4 suggests that the performance of PRSVM is very sensitive to the choice of the phone recognizer. The best dialect identification accuracy in the LDC task is provided by the English-based PRSVM. The baseline AvgEER 7.7 % provided by the English PRSVM is lower than the one reported for the best performing Levantine-based PPRSVM (9.5 %) in [10] on a similar task (note the previously discussed difference in the database setup and that [10] utilized 30 sec. speech segments compared to 11–12 sec. segments used in our study). For the in-house Pan-Arabic corpus, Hindi-based PRSVM provided the best performance across the setups. In order to evaluate the global impact of the different supervector normalizations, independent of the individual phone recognizers, AvgEERs are averaged across all phone recognizers per each strategy (AvgErr). It can be seen that using the sigmoid rather than logarithm in the normalization reduces the classification error in both LDC and in-house Pan-Arabic setups. Incorporating GBF in the sigmoid normalization has on average little or no effect in the LDC task but is helpful in the in-house Pan-Arabic task. Applying hard limiting results in considerable error reduction compared to the original 'GBF+Log' normalization for all PRSVMs and both database scenarios.

Table 5 details EERs per each dialect for English- and Hindi-based PRSVMs, repsectively. The *hard limit* normalization is found to reduce EERs for all dialects compared to the original 'GBF+Log' setup. It can be seen that similarly as in Sec. 3, Egyptian samples are classified in the LDC task with a significantly lower error rate than the other dialects. In the case of the in-house Pan-Arabic task, EERs are well balanced, with the Iraqi dialect yielding slightly lower errors compared to the rest three dialects. Finally, it can be seen that the absence of non-linguistic cues in the in-house Pan-Arabic corpus results in considerable reduction of the classifier performance compared to the LDC task.

Local Bigram Frequency (*LBF*) Normalization	1-vs-3; EER (%)							
	LDC Corpora (Eng. Recognizer)				In-House Pan-Arabic (Hindi Recognizer)			
	GLF	IRQ	LEV	EGY	PS	IRQ	SY	EGY
GBF + Log	10.3	8.2	9.5	2.8	31.8	28.5	33.3	33.0
GBF + Sigmoid	9.5	7.8	8.5	2.6	29.9	26.2	30.5	31.1
Sigmoid	9.6	7.9	8.1	2.5	31.3	27.0	31.7	31.0
Hardlimit	**9.5**	**7.5**	**8.0**	**2.3**	**30.1**	**26.3**	**30.3**	**30.7**

Table 5: Detailed dialect EERs (1-vs-3 task) of English- and Hindi-based PRSVM.

5. Conclusions

This study has analyzed the non-linguistic content of the selected Arabic Conversational Telephone Speech corpora distributed through LDC. It was found that the LDC data sets used in past studies on Arabic dialect identification contain strong non-linguistic cues to the database origin of the recordings. Significant channel differences and distinctive noise characteristics were found in the LDC dialect corpora that are sufficient to perform a relatively successful dialect identification from only silence segments of the recordings. This suggests that normalization of the non-linguistic content may be necessary in order to obtain a fair framework for Arabic dialect identification. In the second part of the study, performance of a PRSVM system on the LDC and in-house Pan-Arabic corpora was evaluated. A simple SVM input supervector normalization utilizing *hard limiting* was shown to consistently reduce dialect identification errors compared to a commonly used normalization by the global N-gram frequency and a logarithmic squashing function.

6. REFERENCES

[1] NIST, "Language recognition evaluation (LRE)," Atlanta, Georgia, Dec. 2011. [Online]. Available: http://nist.gov/itl/iad/mig/lre11.cfm

[2] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. D. Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *INTERSPEECH'02*, Denver, Colorado, 2002, pp. 89–92.

[3] P. Torres-Carrasquillo, E. Singer, W. M. Campbell, T. G. A. McCree, D. A. Reynolds, F. Richardson, W. Shen, and D. E. Sturim, "The MITLL NIST LRE 2007 language recognition system," in *INTERSPEECH'08*, Brisbane, Australia, 2008, pp. 719–722.

[4] 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," in *Odyssey'2010*, Brno, Czech Republic, 2010.

[5] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, Jan. 1996.

[6] M. Zissman, T. Gleason, D. Rekart, and B. Losiewicz, "Automatic dialect identification of extemporaneous conversational, latin american spanish speech," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2, Atlanta, Georgia, May 1996, pp. 777 –780.

[7] H. Suo, M. Li, T. Liu *et al.*, "The design of backend classifiers in PPRLM system for language identification," in *Proc. International Conference on Natural Computation*, Haikou, China, June 2007, p. 678682.

[8] W. Shen, W. Campbell, T. Gleason, D. Reynolds, and E. Singer, "Experiments with lattice-based PPRLM language identification," in *IEEE Odyssey'06: Speaker and Language Recognition Workshop, 2006.*, San Juan, Puerto Rico, June 2006, pp. 1 –6.

[9] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "Phonetic speaker recognition with support vector machines," in *Advances in Neural Information Processing Systems*. MIT Press, 2004, pp. 1377–1384.

[10] M. Akbacak, D. Vergyri, A. Stolcke, N. Scheffer, , and A. Mandal, "Effective Arabic dialect classification using diverse phonotactic models," in *INTERSPEECH'11*, Florence, Italy, 2011.

[11] F. Biadsy, J. Hirschberg, and N. Habash, "Spoken arabic dialect identification using phonotactic modeling," in *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, Athens, Greece, 2009, pp. 53–61.

[12] F. Biadsy, J. Hirschberg, and D. P. W. Ellis, "Dialect and accent recognition using phonetic-segmentation supervectors," in *INTERSPEECH'11*, Florence, Italy, 2011, pp. 745–748.

[13] Y. Lei and J. Hansen, "Dialect classification via text-independent training and testing for arabic, spanish, and chinese," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 85 –96, jan. 2011.

[14] H. Bořil and J. H. L. Hansen, "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1379–1393, August 2010.

[15] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Brno University of Technology, Czech Republic, 2009.

[16] W. Campbell, F. Richardson, and D. Reynolds, "Language recognition with word lattices and support vector machines," in *IEEE ICASSP*, vol. 4, Honolulu, HI, April 2007, pp. 989–992.