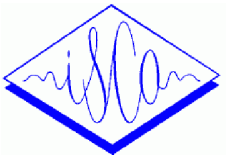
ISCA Archive
http://www.isca-speech.org/archive



Fourth Workshop on Child
Computer Interaction (WOCCI 2014)
Singapore, September 19, 2014

Automatic Assessment of Language Background in Toddlers Through Phonotactic and Pitch Pattern Modeling of Short Vocalizations

Hynek Bořil[1], *Qian Zhang*[1], *Ali Ziaei*[1], *John H. L. Hansen*[1*],
Dongxin Xu[2], *Jill Gilkerson*[2], *Jeffrey A. Richards*[2],
Yiwen Zhang[3], *Xiaojuan Xu*[3], *Hongmei Mao*[3], *Lei Xiao*[3], *Fan Jiang*[3]

[1]Center for Robust Speech Systems (CRSS), University of Texas at Dallas, U.S.A.
[2]LENA Foundation, Boulder, Colorado, USA
[3]Shanghai Children's Medical Center, Shanghai Jiao Tong University School of Medicine,
No. 1678 Dong Fang Road, Shanghai 200127, P.R. China

{hynek,qian.zhang,ali.ziaei,john.hansen}@utdallas.edu
{dongxinxu,jillgilkerson,jeffrichards}@lenafoundation.org
zhangyiwen@hotmail.com

Abstract

This study utilizes phonotactic and pitch pattern modeling for automatic assessment of toddlers' language background from short vocalization segments. The experiments are conducted on audio recordings of twelve 25–31 months old US-born and Shanghainese toddlers. Each recording captures a whole-day sound track of an ordinary day in the toddlers' life spent in their natural environment. In a preliminary study, we observed that in spite of the limited presence of linguistic content in the early age child vocalizations, certain phonotactic and prosodic patterns were correlated with the child's language background. In the current effort, we analyze to what extent these language-salient cues can be leveraged in the context of automatic language background classification. Besides a traditional parallel phone recognition with statistical language modeling (PPRLM) and phone recognition with support vector machines (PRSVM), a novel scheme that utilizes pitch patterns (PPSVM) is proposed. The classification results on very short vocalizations (on average less than 3 seconds long) confirm that both phonotactic and prosodic features capture a language-specific content, reaching equal error rates (EER) of 32.45 % for PRSVM, 31.33 % for PPSVM, and 29.97 % in a fusion of PRSVM and PPSVM systems. The competitive performance of PPSVM suggests that pitch contours carry a significant portion of the language-specific information in toddlers' vocalizations.
Index Terms: language background assessment, toddlers, child vocalization, phonotactic modeling, pitch patterns, PPRLM, PRSVM, PPSVM.

1. Introduction

Thanks to the recent breakthroughs in speech technology, the role of voice interfaces has been gradually extending from an imperfect replacement of a computer keyboard to sophisticated applications in biometrics (user authentication, forensics), education (language learning), and health care (speech-language pathology). While a major part of the research in automatic speech processing has been focused on adult users, recent studies demonstrate its great potential also for children-oriented tasks such as detection of language delay [1], early communication disorders [2], autism [3], computer-aided reading tutoring [4, 5], or emotional state assessment [6]. Other studies focus on automatic assessment of the children vocal development [7] and on boosting the process of early language learning [8].

Our recent study [9] has focused on the analysis of vocalizations from children with American English (*AE*) and Shanghainese (*Shang*) language backgrounds. While the study noted differences in the phonotactic and prosodic domains for the two language backgrounds, it is not clear whether the observed statistical differences are significant and consistent enough to be leveraged in language background discrimination. The main objective of the present study is to design an automatic language background assessment scheme utilizing phonotactics and pitch patterns and verify the significance of the background-specific production differences in a quantitative way. In addition to investigating the role of the two production domains in toddlers' background discrimination, the study aims at advancing the technology for children speech assessment that can benefit future automated child-computer interfaces with applications such as automatic detection of language switching in multi-lingual children or language acquisition assessment.

State-of-the-art language recognition systems for adult speech, as seen in recent National Institute of Standards and Technology Language Recognition Evaluation (NIST-LRE) [10] submissions, typically utilize one or a combination of several of the following strategies: cepstral coefficients with shifted delta cepstra (SDC) [11], Gaussian mixture modeling with universal background models (GMM-UBM) and GMM supervectors [12] and i-vectors [13], phonotactic models realized by parallel phone recognizers and language modeling (PPRLM) [14], and phone recognizers combined with support vector machines (PRSVM) [15, 16].

In our study, PPRLM and PRSVM systems are used for

*This project was funded by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

phonotactic-based language background assessment. In both PPRLM and PRSVM, speech signal is first decoded by a phone recognizer or a set of phone recognizers, respectively, into a sequence of phones [15] or phone lattices [16]. A phone recognizer (PR) can be trained on a language or a mixture of languages that may be unrelated to the target languages [17]. The idea is that when decoding an utterance, even from an unrelated language, the PR will work as an acoustic event detector and generate sequence of phones or phone lattices that reflect the PR's acoustic model states closest to the processed signal. It is expected that different input language will produce different PR outputs. In PPRLM, statistical *N*-gram language model (LM) statistics are trained on output phone sequences from multiple phone recognizers. Evaluation stage combines PR and LM decoding. In PRSVM, PR outputs are typically normalized, processed by a squashing function, stacked into supervectors, and passed to SVM classifiers.

To incorporate longer-term pitch contours in the automatic language background classification, this study proposes a so called pitch pattern SVM (PPSVM) scheme where a pitch pattern tokenizer is used to generate sequences of discrete symbols – pitch pattern words – which are then modeled by SVM in a similar manner as seen in conventional PRSVM. The remainder of the paper is organized as follows. First, the corpus of children recordings is presented. Second, phonotactic and pitch pattern modeling techniques are discussed. Finally, evaluation results are discussed.

2. Corpus of US/Shanghainese Toddlers

The corpus used in this study represents a subset of the recordings introduced in [9] and contains six US-born and six Shanghainese children of ages spanning 25–31 months. The subjects were selected to have identical age and gender distribution in both language groups (age in months followed by gender; *M* – male, *F* – female): 25 F, 26 F, 2 x 27 F, 27 M, and 31 F. The recordings were collected in the children's natural environment using a lightweight digital recorder placed in the pocket of the subject's clothes [7]. The recordings represent whole day sessions (typically 14–16 hours).

The sessions contain multiple environments (home, car, shopping malls and restaurants, grandparent visits) and activities (playing, eating, taking a rest) over time. Besides child vocalizations, the recordings capture all ambient sounds and noises occurring in the vicinity of the subject. Due to the varying environment and presence of nonstationary background noises and speech from secondary speakers in the recordings,

Set	American English (*AE*)				Shanghainese (*Shang*)			
	#Subj.	#Samp.	Avg Dur. /σ (sec)	Total Dur. (mins)	#Subj.	#Samp.	Avg Dur. /σ (sec)	Total Dur. (mins)
Train	6 (1M, 5F)	2193	2.2 (1.9)	79.7	6 (1M, 5F)	1767	2.8 (4.9)	83.5
Test		1096	2.3 (2.6)	41.7		883	2.7 (3.0)	39.9

Table 1: Corpus content; *#Subj.* – number of subjects (*M/F* – males/females); *#Samp.* - number of vocalization samples; *Avg. Dur.* – average sample duration (and its standard deviation); *Total Dur.* – total duration of set samples in minutes.

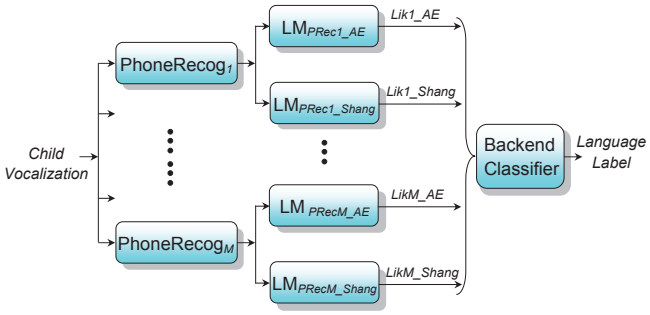


Figure 1: Parallel phone recognition and statistical language modeling (PPRLM); *AE* – American English (US-born subjects), *Shang* – Shanghainese.

it is difficult to reliably locate child vocalization segments through automated processing. For this reason, time boundaries of child vocalizations were manually labeled for each session by human annotators. Following the data disclosure agreement, no additional information besides the time labels was collected in the labeling process. A minimum of 20 minutes of vocalizations were labeled per each subject. More details on the process can be find in [9]. For the language background assessment experiments, 2/3 of the vocalization samples per subject were assigned to the training set and 1/3 to the evaluation set. Due to the limited number of subjects available, we have chosen to form closed-subject sets to preserve as much speaker variability as possible. In the closed-subject scenario, all subjects appear in both the training and evaluation sets (with different vocalization samples being used for training and evaluation). The corpus statistics are detailed in Table 1.

3. Phonotactic Models

This section presents two approaches to phonotactic modeling that utilize a bank of parallel phone recognizers and statistical language models (PPRLM) and phone recognizers combined with support vector machines (PRSVM).

3.1. PPRLM

PPRLMs are frequently used to produce front-end features in speaker and language identification systems for adult speech [18, 19, 15]. The structure of a PPRLM system used in the current study is depicted in Fig. 1. Incoming speech segments are fed to a bank of phone recognizers whose acoustic models were trained on a variety of languages. During decoding, each incoming segment is assigned a label representing the acoustically closest phone model. This process can be viewed as tokenization – conversion of a continuous acoustic signal into a sequence of discrete symbols. Due to acoustic-phonetic differences across languages, a combination of multiple phone recognizers trained on different languages can be expected to provide more detailed tokenization. A mismatch between the processed language and the language of the recognizer is often beneficial in the phonotactic modeling [20] and as demonstrated in [9], phone recognizers trained on adult speech are in fact sensitive to the variability in child vocalizations.

In the training stage of PPRLM, a separate statistical language model (LM) is trained for each phone recognizer and tar-

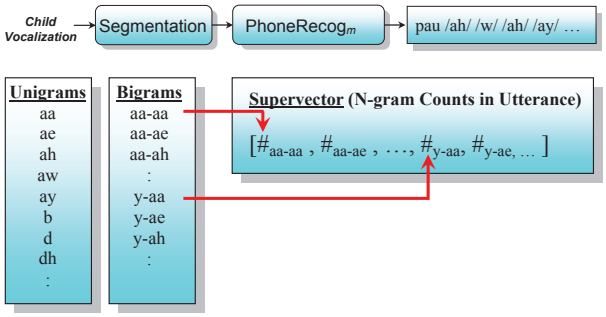


Figure 2: Phone recognition combined with support vector machines (PRSVM); upper part – extraction of phone sequence from phone recognizer; bottom part – construction of a vocalization-level bigram supervector.

get language (*AE*, *Shanghainese*). For example in the case of a Hungarian phone recognizer, *Shanghainese* trainining samples are decoded by the recognizer and the output phone sequences are used to calculate the *N*-gram statistics for the LM_{PRecHu_Shang}. The same is repeated for *AE* training samples to extract LM_{PRecHu_AE}.

In the decoding stage, the vocalization segment is processed by all PPRLM branches, generating a vector of likelihoods that are subsequently passed to the backend classifier. In our case, unweighted sums of the *AE* and *Shang* likelihoods, respectively, are calculated and compared to decide the resulting language label. A set of nine BUT (Brno University of Technology) phone recognizers [21] is used in this study: English, Czech, Hungarian, Russian, German, Hindi, Japanese, Mandarin, and Spanish.

3.2. PRSVM

The PRSVM used in this study follows [15, 17]. The same set of 9 phone recognizers as in the previous section is used also here. However, unlike in PPRLM, here each phone recognizer constitutes an autonomous PRSVM system. Each input vocalization sample is first phone-decoded and subsequently, frequencies of all *N*-grams appearing in the output phone sequence are calculated, normalized [15, 17, 20], processed by a so called squashing function [22, 23], and stacked into a supervector where each dimension represents a normalized frequency of a particular *N*-gram (see Fig. 2). Bigram language models are used in all our PRSVM setups. Four squashing functions are considered in our experiments: log, square root, hard limit, and sigmoid [20]. Following [24, 20], a MAP-adaptation of a universal *N*-gram language model is used to extract the final form of the complete supervector. Finally, dimensionality reduction through frequency- or salience-based *N*-gram selection is applied [20]. The supervectors extracted from the training vocalizations are used to train binary SVM classifiers to distinguish the two target language backgrounds.

4. Pitch Pattern SVM (PPSVM)

Here, we utilize a simple automatic pitch pattern extraction technique established in [9] and inspired by [25], and propose its use in a PRSVM-like processing scheme. WaveSurfer [26] is first used to extract a pitch contour from each child vocalization segment. Subsequently, voiced sections of continuous nonzero F_0 values are identified. Each section is passed through a median filter and processed by a regression analysis performed on a sliding window of the length T_{win} shifted with the step T_{step}. Voiced sections shorter than T_{win} are excluded from the process. A straight line is fit into the F_0 contour within the window by means of linear regression. If the slope of the regression line is steep enough to cross a frequency band F_{th} within the range of the window, the segment is assigned a rising/falling pattern element; otherwise a flat pattern is assigned. Figure 3 shows an example of the pattern matching process. The original regression lines are presented as solid lines, the slope of the dashed lines suggests the decided pattern (up/flat/down).

While there are typically 30–60 different phones in a language, the elementary pitch patterns have only three variants (up, flat, down). In order to model a longer temporal context and increase the pitch pattern variability in the supervector processing scheme, short sequences of adjacent patterns are grouped together to form pattern 'words'. For example, assuming trigram words, a sequence {up, flat, flat, down, down, up} can be transformed to two trigrams {up-flat-flat, down-down-up} given there is no overlap between the two words. We denote this as *3gram*$_{Step3}$ – 3-gram words formed with the skip-rate of three. *3gram*$_{Step2}$ denotes 3-gram words with the skip-rate of two, i.e., adjacent words will share one pattern in common: {up-flat-flat, flat-flat-down, flat-down-down, down-down-up}. The pitch pattern word sequences are subsequently transformed into *N*-gram supervectors in the same way as the phone patterns in the previous section (applying bigram processing on the 'words'), yielding a PPSVM system. One might argue that the two-stage *N*-gram processing is redundant, however, its benefits are in a finer control of the supervector content through the arbitrary skip rate.

5. Evaluations

In the first step, individual PRSVM and PPSVM systems utilizing only a single tokenizer (single phone recognizer or single pitch pattern 'word' setup) were evaluated. Table 2 presents two best performing setups per each system. The language background classification performance is evaluated by means of equal error rate (EER), which represents an operating point of the binary classifier where the percentage of *AE* being mistakenly classified as *Shang* and vice versa are balanced. EER has been popular in pairwise language identification evaluations in the NIST LRE campaign [10] and unlike accuracy,

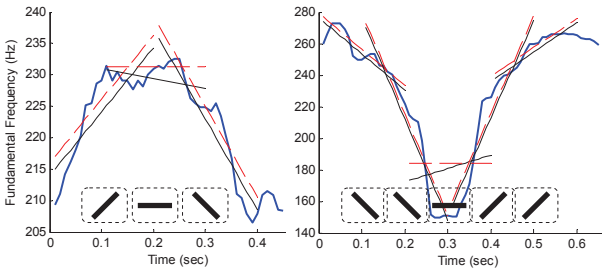


Figure 3: Example of pitch pattern extraction: window length T_{win} = 200 ms, window step $T_{step} = T_{win}/2$, pattern clustering threshold $F_{th} = 15$ Hz.

System	Tokenizer	Squashing Function	Dimension Reduction	EER (%)
Phone Recognition SVM (PRSVM)	Hindi	Sqrt	Frequency	37.49
	Mandarin	Hard Limit	Salience	37.70
Pitch Pattern SVM (PPSVM)	$2gram_{Step1}$	Sigmoid	Salience	39.62
	$3gram_{Step1}$	Sigmoid	Salience	39.62

Table 2: Performance of individual PRSVM and PPSVM systems (single tokenizer per system).

is unbiased by different class sample sizes. Another advantage of EER is its easy interpretability in terms of system performance as an operating point on the detection error tradeoff (DET) curve [27].

It can be seen that the top PRSVM systems and PPSVM systems provide comparable performance. It is noted that while these EERs may seem relatively high, the classification is conducted on very short child vocalizations (shorter than 3 sec on average) that often contain babbling rather than a linguistic content intelligible to adults.

In the next step, a fusion of systems within their respective domain was evaluated (see Table 3). Here, the EER is noticeably reduced for both PRSVM and PPSVM systems and in this case, a fusion of PPSVMs provides superior performance with 31.33 % EER. The good performance of the pitch pattern based systems is somewhat surprising. In [9], bigram pitch pattern statistics did not reveal any significant differences between *AE* and *Shang* vocalizations in spite of the latter being a tonal language. In this sense, the PPSVM clearly benefits from longer context models (bigram models of 3-gram pitch pattern words). PPRLM reaches comparable results to the single-tokenizer systems in Table 2 but lags significantly behind the fused PRSVM and PPSVM.

Finally, performance of three top-ranking PRSVM and PPSVM fused systems is presented in Table 4. The fusion provides a further absolute EER reduction by 1.36 %, resulting in 29.97 % EER. It can be seen that the PPSVM systems found in the top two fusions here are identical with those in Table 3. On the other hand, different combinations of PRSVMs were found optimal in a with-PRSVM domain fusion versus PRSVM–PPSVM fusion.

System	Fusion Components	Squashing Function	EER (%)
Parallel Phone Recognition Statistical LM (PPRLM)	EN,CZ,RU,GER,HIN, JAP,MAN,SPA	N/A	38.75
Phone Recognition SVM (PRSVM)	HU,RU,GER,JAP,MAN,SPA	Sqrt	32.45
	HU,RU,GER,HIN,JAP,MAN,SPA	Sqrt	32.64
Pitch Pattern SVM (PPSVM)	$3gram_{Step2}$, $2gram_{Step1}$, $1gram_{Step1}$	Sqrt	31.33
	$3gram_{Step2}$, $2gram_{Step2}$	Sqrt	31.33

Table 3: Performance of domain-constrained fusion of PPRLM, PRSVM, and PPSVM systems.

System	Fusion Components	Squashing Function	EER (%)
Phone Recognition & Pitch Pattern SVM (PRPPSVM)	$3gram_{Step2}$, $2gram_{Step2}$, HU,GER,HIN,JAP,MAN,SPA	Sqrt	29.97
	$3gram_{Step2}$, $2gram_{Step2}$, $1gram_{Step1}$, HU,GER,HIN,JAP,MAN,SPA	Sqrt	29.97
	$3gram_{Step2}$, $3gram_{Step1}$, $2gram_{Step2}$, $3gram_{Step2}$, HU,RU,GER,HIN,JAP,MAN,SPA	Sqrt	30.12

Table 4: Performance of across-domain fusion of PRSVM and PPSVM systems.

Studies on the role of prosody in spoken language acquisition suggest that young children strongly rely on prosody in their comprehension and production of spoken language [28]. Before acquiring grammar and comprehending utterances, they have to be able to locate linguistically relevant segments in speech stream. Prosodic changes (e.g., pauses, lengthening of syllables, and pitch resetting) are often related to linguistic boundaries and past studies were able to measure correlation between such events and the level of speech comprehension in infants and young children [29]. In this sense, due to their extensive focus on prosody in the early stages of language acquisition, toddlers are likely to reproduce prosodic patterns heard from their adult peers, even prior to being able to produce a 'valid' phonetic content. This might be an explanation of the comparable or even superior performance of the pitch pattern based PPSVM on the toddlers' vocalizations in our study compared to the phonotactic PRSVM and PPRLM systems.

6. Conclusions

This study investigated the viability of automatic language background assessment from short segments of toddlers' vocalizations via phonotactic and pitch pattern based classifiers. Considering the limited presence of linguistic content in the vocalizations and the very short segment durations (on average less than 3 seconds long), the classification results seem quite encouraging. In the within-domain system fusion, the newly proposed pitch pattern-based systems slightly outperformed phonotactic systems, which suggests that prosody carries strong language background cues in Shanghainese and US-born toddlers. This might seem intuitive as one of the compared languages is tonal and the other is not. However, our previous study conducted on pitch pattern bigrams did not reveal any significant differences. Hence, the good performance of the PPSVM systems here has to be attributed to the longer context models (6-grams) where the two language backgrounds start to exhibit more significant differences. The prevailing presence of language background cues in the pitch contours rather than phonotactics can be related to the prominent role of prosody in spoken language acquisition in infants and young children. Based on this observation, the authors hypothesize that in the initial stages of language acquisition, when toddlers focus mainly on adult speech prosody, the language-related content of their vocalizations is likely to be dominated by prosodic rather than phonetic cues.

7. REFERENCES

[1] D. Xu, U. Yapanel, S. Gray, J. Gilkerson, J. Richards, and J. H. L. Hansen, "Signal processing for young child speech language development," in *1st Workshop on Child, Computer, and Interaction*, Chania, Greece, Oct. 2008.

[2] P. Zlatník and R. Čmejla, "Disordered speech assesment using different speech parameterizations," in *19th International Congress on Acoustics*, Madrid, Spain, 2007, pp. 1–4.

[3] D. Xu, J. Gilkerson, J. Richards, U. Yapanel, and S. Gray, "Child vocalization composition as discriminant information for automatic autism detection," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, 2009, pp. 2518–2522.

[4] J. Tepperman, J. Silva, A. Kazemzadeh, H. You, S. Lee, A. Alwan, and S. Narayanan, "Pronunciation verification of children's speech for automatic literacy assessment," in *Proc. ICSLP'06*, Pittsburgh, PA, USA, 2006, pp. 845–848.

[5] A. Hagen, B. Pellom, and R. Cole, "Highly accurate children's speech recognition for interactive reading tutors using subword units," *Speech Comm.*, vol. 49, no. 12, pp. 861–873, 2007.

[6] C. L. S. Yildirim, S. Lee, A. Potamianos, and S. Narayanan, "Detecting politeness and frustration state of a child in a conversational computer game," in *Proc. EUROSPEECH'05*, Lisbon, Portugal, 2005, pp. 2209–2212.

[7] D. Xu, J. Gilkerson, and J. A. Richards, "Objective child vocal development measurement with naturalistic daylong audio recording," in *INTERSPEECH'12*, 2012.

[8] P. K. Kuhl, B. T. Conboy, S. Coffey-Corina, D. Padden, M. Rivera-Gaxiola, and T. Nelson, "Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e)," *Philos. Trans. R. Soc.*, vol. B, no. 363, pp. 979–1000, 2008.

[9] H. Bořil, Q. Zhang, P. Angkititrakul, J. H. L. Hansen, D. Xu, J. Gilkerson, and J. A. Richards, "A preliminary study of child vocalization on a parallel corpus of US and Shanghainese toddlers," in *INTERSPEECH'13*, Vancouver, Canada, 2013, pp. 2405–2409.

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

[11] 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, 2008, pp. 719–722.

[12] E. Singer, P. Torres-Carrasquillo, D. Reynolds, A. McCree, F. Richardson, N. Dehak, and D. Sturim, "The MITLL NIST LRE 2011 language recognition system," in *Proc. IEEE ICASSP'12*, June 2012, pp. 209–215.

[13] D. R. R. D. N. Dehak, P. A. Torres-Carrasquillo, in *Language Recognition via Ivectors and Dimensionality Reduction*, Florence, Italy, Aug. 2011, pp. 857–860.

[14] 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, June 2006, pp. 1 –6.

[15] 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.

[16] W. Campbell, F. Richardson, and D. Reynolds, "Language recognition with word lattices and support vector machines," in *Proc. IEEE ICASSP'07*, vol. 4, april 2007, pp. 989–992.

[17] 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.

[18] 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.

[19] 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.

[20] Q. Zhang, H. Bořil, and J. H. L. Hansen, "Supervector pre-processing for PRSVM-based Chinese and Arabic dialect identification," in *Proc. IEEE ICASSP'13*, Vancouver, May 2013, pp. 7363–7367.

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

[22] W. Campbell, F. Richardson, and D. Reynolds, "Language recognition with word lattices and support vector machines," in *Proc. IEEE ICASSP'07*, vol. 4, Honolulu, HI, April 2007, pp. 989–992.

[23] H. Bořil, A. Sangwan, and J. H. L. Hansen, "Arabic dialect identification - 'Is the secret in the silence?' and other observations," in *INTERSPEECH'12*, Portland, Oregon, 2012.

[24] B. Xu, Y. Song, and L. Dai, "The adaptation schemes in PR-SVM based language recognition," in *Chinese Spoken Language Processing, 2008. ISCSLP '08. 6th International Symposium on*, dec. 2008, pp. 1 –4.

[25] R. D. Kent and A. D. Murray, "Acoustic features of infant vocalic utterances at 3, 6, and 9 months," *The Journal of the Acoust. Soc. of Am.*, vol. 72, no. 2, pp. 353–365, 1982.

[26] K. Sjolander and J. Beskow, "WaveSurfer – an open source speech tool," in *Proc. of ICSLP'00*, vol. 4, Beijing, China, 2000, pp. 464–467.

[27] A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, and M. A. Przybocki, "The DET curve in assessment of detection task performance," in *EUROSPEECH'97*, Rhodes, Greece, 1997, pp. 1895–1898.

[28] S. R. Speer and K. Ito, "Prosody in first language acquisition - acquiring intonation as a tool to organize information in conversation," *Language and Linguistics Compass*, vol. 3, no. 1, pp. 90–110, 2009.

[29] L. Gerken, "Prosody's role in language acquisition and adult parsing," *Journal of Psycholinguistic Research*, vol. 25, no. 2, pp. 345–356, 1996.