A Preliminary Study of Child Vocalization on a Parallel Corpus of US and Shanghainese Toddlers

Hynek Bořil[1], *Qian Zhang*[1], *Pongtep Angkititrakul*[1], *John H. L. Hansen*[1*],
Dongxin Xu[2], *Jill Gilkerson*[2], *Jeffrey A. Richards*[2]

[1]Center for Robust Speech Systems (CRSS), University of Texas at Dallas, U.S.A.
[2]LENA Foundation, Boulder, Colorado, USA
{hynek,qian.zhang,pongtep.angkititrakul,john.hansen}@utdallas.edu

Abstract

This paper studies various aspects of child vocalization as captured in a newly established parallel corpus of sixteen 18–31 months old US and Shanghainese toddlers. The recordings were acquired in 16-hour sessions during an 'ordinary' day in the child's natural environment and manually labeled. The vocalization characteristics are studied by means of phonotactic and prosodic analysis with emphasis on automatic processing. In the phonotactic domain, a Gaussian mixture model (GMM) tokenizer, a bank of phone recognizers, and formant tracking are used to analyze the movements in the acoustic-phonetic space. In the prosodic domain, pitch patterns, duration, and rhythm are analyzed. Besides strong individual-specific characteristics of the subjects in some of the domains considered, the two language groups show differences in the occupation of the F_1–F_2 formant space, choice of pitch pattern durations, and consistency in producing complex phonetic patterns.

Index Terms: children vocalization, speech acquisition, phonotactic modeling, pitch patterns, rhythmicity parameters

1. Introduction

While the mechanisms of speech perception, production, and language acquisition in children have been studied for an extensive period of time, it is just recently where speech technology has matured to the level to be able to considerably contribute to these domains. Recent studies have shown the potential of automatic speech processing to perform tasks such as detection of language delay [1], early communication disorders [2], and autism [3], computer-aided reading tutoring [4], or emotional state assessment [5]. While most current speech systems are designed with adult users in mind, better understanding of children speech perception and production can benefit the design of more effective children-oriented engines [6].

Observing the importance of social interactions and their impact on early language learning [7], the current studies focus on the automatic assessment of the child vocal development [8] with the prospect of providing the child's peers with constructive feedback. Our study aims at expanding the ensemble of techniques for automatic child vocalization assessment presented in [8]. The proposed methods are applied on the newly established corpus of US and Shanghainese children across an age range of 18–31 months. The remainder of the paper is organized as follows. First, wet present the corpus of children recordings. The next part is dedicated to prosodic analyses utilizing pitch patterns and speech rhythm measure. The third part discusses phonotactic techniques.

*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.

2. Corpus of US/Shanghainese Toddlers

The corpus utilized in this study captures audio recordings of eight US and eight Shanghainese children across an age range of 18–31 months (five female and three male subjects per each language background) and represents a subset of recordings from an ongoing large scale parallel US and Shanghainese children speech acquisition campaign. The subjects in our corpus were selected to have, in terms of age and gender, an identical counterpart in each language group (age in months/gender): 18 M, 23 M, 25 F, 26 F, 2 x 27 F, 27 M, and 31 F. The recordings were acquired in the children's natural home environment using a lightweight digital recorder [8]. The recorder was placed in the pocket of the child's clothes, allowing for free and natural movement. For each subject, a whole day recording was acquired (typically 14–16 hours).

The sessions capture segments of the child staying at home (playing, eating, taking a rest), accompanying parents to shopping malls and restaurants, visiting grandparents, etc. Due to the nature of the recordings, the audio tracks contain not only vocalizations of the child subject but also voices of siblings and peers as well as other ambient sounds (TV, radio, kitchen sounds, barking dogs, sounds produced by friction of the clothes and recording device, etc.). The level of these *secondary* sounds varies and in some instances reaches or overcomes the child's voice. At times, identifying the instances of the child's vocalization may become difficult, especially when the subject interacts with similar aged siblings or friends, or while listening to children TV programs.

The time instances of the child vocalization (CV) were manually annotated by two labelers. For each subject, approximately (but never less than) 20 minutes of vocalized segments were selected. The segments were chosen as follows: CV segments without interfering secondary vocalizations were picked; CV boundaries were set to exclude barge-ins. If a barge-in occurred during the first word or word-like sound of the CV, the segment was dropped. In other cases, a CV representing even only a part of speech would be kept. CVs were dropped when interference with sounds of a harmonic structure (music, squeaking toys) occurred. Following these criteria, approx. 200–800 segments per session were identified to reach the 20 minutes of child vocalization per subject; however, 400–700 segments were predominant for most sessions. The subject (US, 23 M) with the record number of 843 CV segments preferred to

Copyright © 2013 ISCA

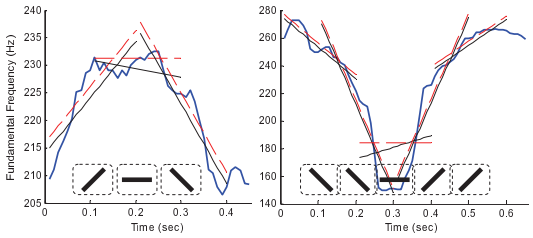

Figure 1: Extraction of pitch patterns; example: T_{win} = 200 ms, $T_{step} = T_{win}/2$, $F_{th} = 15$ Hz.

utter single syllable words *up* and *juice* at numerous instances. On the other hand, CV segments of more extensive length usually contained a higher degree of babbling [9]. Surprisingly, even the oldest and/or most articulate subjects would sometimes return to babbling, especially when playing alone. In total, 4326 CV segments were labeled for the US group and 3288 for the Shanghainese group to meet the target quota of 20 mins/subject.

3. Pitch Patterns

In [10], seven simple shapes were manually fit into the F_0 contours and used to analyze the frequency of repetition of pitch patterns in infant vocalic utterances. In this paper, we utilize a simple automatic pitch pattern production/analysis technique inspired by [10]. In the first step, WaveSurfer [11] is used to extract the pitch track from each CV segment. Subsequently, *voiced islands* of continuous nonzero F_0 values are found. Each island is median filtered and processed by a sliding window of length T_{win} shifted with a step T_{step}. Voiced islands shorter than T_{win} are dropped from the analysis. A straight line is fit into the F_0 values captured by the window by means of linear regression. If the regression line is steep enough to cross a frequency band F_{th} within the range of the window, the window segment is assigned a rising/falling pattern element; otherwise a flat pattern is assigned. Figure 1 demonstrates the pattern matching process. The regression lines are presented as solid lines, the slope of the dashed lines denotes the decision pattern. The actual slope value is subsequently discarded and only the direction is kept (rising/flat/falling). Finally, frequency of pattern unigrams and their sequences (N-grams) is calculated.

For several settings of T_{win} in the range of 50–200 ms (and F_{th} 5–20 Hz), similar patterns would prominently appear for both US (*AE*) and Shanghainese (*Shang*) groups. Fig. 2 shows pattern bigram frequencies for *AE* subjects and Table 1 details both unigram and bigram frequencies for *AE* and *Shang* groups. It can be seen that the falling pattern is dominant, followed by the rising and flat pattern. The same is replicated in bigrams by the double-falling, double-rising and double-flat patterns. The

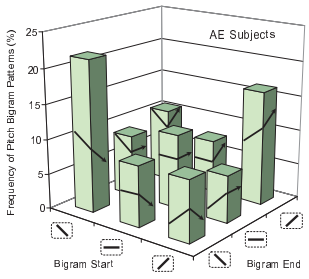

Figure 2: Frequency of bigram patterns in AE subjects; T_{win} = 50 ms, $T_{step} = T_{win}$, $F_{th} = 5$ Hz.

order of preference of the subsequent bigrams is slightly different for *AE* and *Shang*, however, the frequencies are similar. Given that Shanghainese is a tonal language, one might expect the differences in pattern choices would be more prominent. However, for the range of settings tested by the authors, such a hypothesis was not confirmed. Finally, Fig. 3 presents the frequency of appearance of pitch patterns with respect to their length (e.g., the leftmost plots represent CV segments that contain only one pitch pattern element). For each length, *AE* (left) and *Shang* (right) are presented. Here and in the rest of the text, the boxplots read as follows: the edges of the box represent 25^{th} and 75^{th} percentiles, the central mark is the median, the whiskers extend to the most extreme points that are not considered outliers, and the outliers are plotted as individual points. Fig. 3 suggests that the *AE* group produces unigram and bigram patterns almost equally likely, higher order N-grams being in decline, while *Shang* prefers bigrams and trigrams to unigram patterns. *Shang* still chooses four-grams more frequently than *AE*. In other words, in most CV segments, *Shang* tends to produce consistently longer pitch patterns than *AE*. This correlates well with the higher number of CV segments required for the *AE* to fulfill the same 20 mins/subject requirement (Sec. 2).

4. Speech Rate

This section utilizes a speech rate estimation algorithm implemented following [12]. The algorithm derives a sequence of prominent minima and maxima in a smoothed RMS envelope of the acoustic signal. The algorithm adaptively sets the decision thresholds to identify the envelope extremes and filters out groups of extremes that are likely related to a single syllable nucleus.In our study, the inverse of the distance between the established envelope extremes is used to estimate the rhythmicity in the vocalization segments – for simplicity denoted *speech rate*. Fig. 4 shows the speech rate distributions for all 16 children. It can be seen that the rates vary across the individuals and do not display any obvious trends with age. In a separate analysis of the overall rhythm distributions for the two language groups, the 25^{th}–75^{th} percentile boxplots displayed a nearly perfect overlap, suggesting there are no observable effects of the language factor on the speech rhythm in our dataset.

5. Formant Analysis

Location of formants during phonation in children reflects both physiological characteristics of their vocal tracts as well as the content and variability of phonation. In [13], it was shown that the first formant F_1 continues to decrease until approx. 30 months of age while F_2 was found steadier starting from 18 months. In [14], $F_{1,2}$ stayed relatively unchanged in 15–24 months, followed by a significant decrease in 24–36 months. In our study, the first two formants are extracted using WaveSurfer. Formant tracks in the *voiced islands*, established from pitch analysis are smoothed by a moving average filter, with unvoiced segments discarded. The distribution of formants in the F_1–F_2 plane is shown in Fig. 5, 6 by means of 2–σ ellipses [15]. The left hand side of Fig. 5 shows an example scatter plot of formant realizations for *AE* subject 23 M. It can be assumed

AE	40.3	34.3	25.4	21.7	16.4	10.7	10.6	8.8	8.8	8.7	7.6	6.8
Shang	41.5	31.1	27.4	23.1	14.8	12.6	8.8	8.7	8.3	9.3	7.2	7.1

Table 1: Frequency of pitch patterns (%).

that some of the leftmost F_1 values result from errors in formant tracking, however, most values occupy a meaningful interval. The right hand part compares 2–σ ellipses for the complete *AE* and *Shang* groups. The tilt of the *Shang* distribution is steeper as a result of F_1 samples occupying a more compact space than its *AE* counterpart. Fig. 6 details distributions for ten female subjects. While the overall orientation of the ellipses follows that from Fig. 5, there is also a notable reduction in the ellipse surface for both *AE* and *Shang* subjects 31 F compared to their younger colleagues. As a matter of fact, the ellipse surfaces were found to mostly reduce with increasing age of the female subjects (*AE* $3.17, 3.16, 2.94, 2.91, 2.51(\times 10^6)$; *Shang* $3.01, 3.01, 3.102.80, 2.36(\times 10^6)$). On the other hand, the surfaces in male subjects did not follow such a trend (consistently increasing in *AE* males and being rather steady in *Shang*). Clearly, the available amount of subjects per page group in this study does not allow any significant analyses of the age effects. However, the reduction observed in females seem to follow those in [14].

6. GMM Tokenizer

In this section, the acoustic-phonetic space in the child vocalization segments is studied by means of a Gaussian mixture model (GMM) tokenizer. The concept of a GMM tokenizer has been popular in the field of speaker and language identification [16]. In our study, the GMM is trained using expectation-maximization algorithm on the complete set of CVs from all subjects. Thirteen static mel frequency cepstral coefficients (MFCC) are extracted from 25 ms window shifted with a step of 10 ms. The GMM mixtures model the acoustic space occupied by the subjects during phonation. Since the MFCC features carry both speaker-dependent and linguistic content-dependent characteristics, we set the number of GMM mixtures intentionally low (32). In this case, the mixtures are forced to be shared between the subjects and hence, profile more towards the phone-like groups rather than towards the speaker specific characteristics. After the GMM is trained, the mixtures are split to form individual single Gaussian states in single state hidden Markov models (HMM). All subject CVs are subsequently decoded by the HMMs, generating a sequence of symbols (indices of the original GMM mixtures). The symbol sequences reflect the transitions of the individual's phonation in the acoustic-phonetic space. We opted for using 'mixture' HMMs rather than Gaussian models for decoding to be able to regulate the state transition probabilities and hence, the rate of transitions between the 'mixture states'. The output strings from the tokenizer are processed by means of sequential pattern analysis.

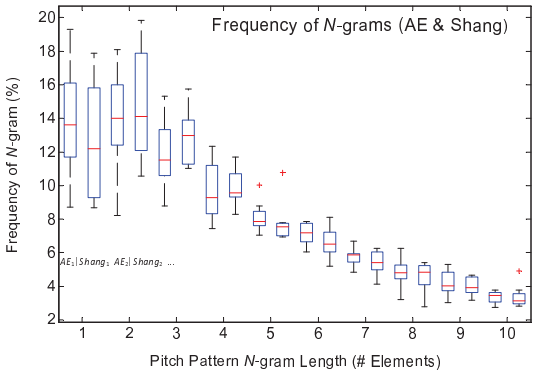

Figure 3: Frequency of *N*-gram occurrences.

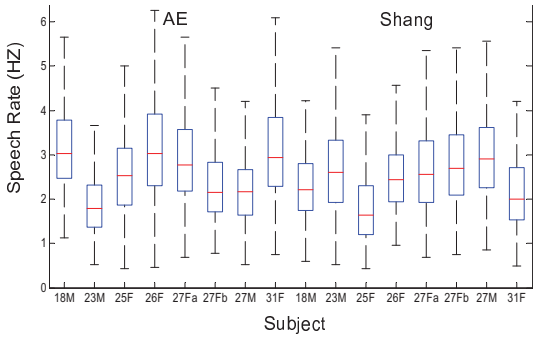

Figure 4: Speech rate distribution in *AE* and *Shang* subjects.

6.1. Sequential Pattern Analysis

Sequential pattern mining [17] is employed to discover frequent sub-sequences as common patterns shared among the subjects. The motivation of this analysis is that frequent sequential patterns, which reflect strong associations within the test group, capture the underlying combination of acoustic units produced by the children. Our assumption is that longer unique sequential patterns are a sign of higher language proficiency. In our study, we employed a PrefixSpan approach [18] to mine the hidden sequential patterns in the CVs. The PrefixSpan approach is based on a divide-and-conquer framework, where the first scan is performed to derive the set of 1-symbol long sequential patterns of all data. Subsequently, each sequential pattern is treated as a prefix and the complete set of sequential patterns can be partitioned into different subsets according to different prefixes. Finally, PrefixSpan is set to find all the sub-sequences whose occurrence frequency in the set of sequences is no less than the minimum support threshold. More details regarding PrefixSpan technique can be found in [18].

6.2. Sequential Patterns of GMM mixtures

A consecutive sequence of the same decoded mixtures are grouped into a single unit (i.e., 1 1 1 1 3 3 4 4 4 4 4 $\rightarrow$ 1 3 4). We are interested in analyzing sequential patterns of the GMM mixtures (i.e., the transition patterns between GMM mixtures in the acoustic space) consistently shared by the members of the two language groups. The minimum support threshold was set to 1 % of the total number of frame clusters. The analysis identified that all 32 mixtures were actively involved in the patterns of both *AE* and *Shang* subjects. In addition, there were

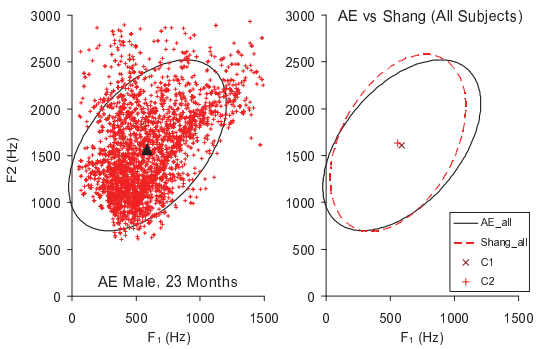

Figure 5: Left – example of F_1–F_2 realizations in a subject; right – comparison of the effect of language background on the formant space.

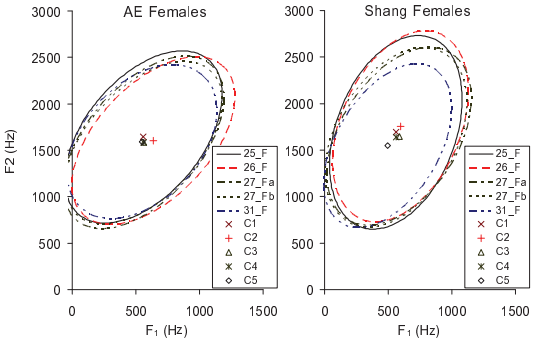

Figure 6: 2–σ ellipses for female *AE* and *Shang* subjects.

586 unique two-symbol patterns, 890 3-symbol patterns, 361 4-symbol patterns, and 52 5-symbol patterns that appeared more frequently in the *Shang* CV segments than the threshold. A formatted version of this result is presented as 32/586/890/361/52. Using the same notation, the *AE* subjects were able to produce 32/472/210/0/0 unique patterns. This means that besides the 1-symbol, *Shang* subjects generated more unique mixture transition patterns than their *AE* counterparts for all sequential lengths. For illustration, Fig. 7 presents the occurrence of 2-symbol patterns for the two language groups, where each pixel represents a transition pattern between the row index mixture and the column index mixture, and the brighter the color, the more transition occurrences. As can be seen, *Shang* phonation results in approx. 10 % higher 2-symbol transition coverage than the *AE* group.

7. Parallel Phone Recognizers

Parallel phone recognizers (PPR) are frequently used as a front-end in speaker and language identification systems [19, 20, 21]. Similar to the GMM tokenizer, the PPR generates strings of symbols that can be subsequently modeled and classified. Unlike the case for our GMM tokenizer, the phone recognizers are frequently trained on a material seemingly unrelated to the target task. For example, in language/dialect identification, phone recognizers trained on unrelated languages often provide more relevant information than matched recognizers [22]. The reason for that is that phone models trained on various languages are sensitive to different aspects of acoustic variation and, when combined, can provide substantially finer resolution of the acoustic-phonetic nuances compared to the recognizer trained on the target domain. This concept is explored in this section by utilizing recognizers trained on non-English and non-Shanghainese subjects. In addition, the recognizers were trained on adult speech which introduces yet another, and most likely, the most substantial mismatch. While this setup may seem completely counterintuitive, as far as the recognizers are

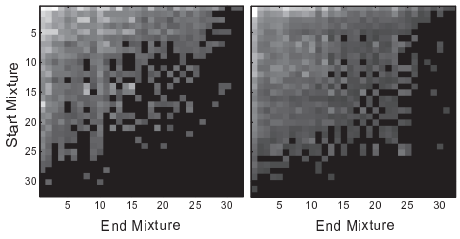

Figure 7: Mixture transitions in *AE* (left) and *Shang* (right) subjects.

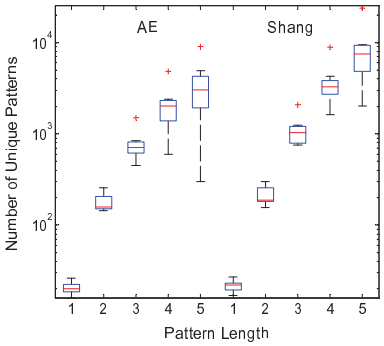

Figure 8: Number of exclusive phone-like patterns in *AE* and *Shang* group. Boxes – pattern counts across 7 BUT phone recognizers.

capable of producing non-trivial responses to the input CV segments, the concept should be meaningful. A set of seven BUT (Brno University of Technology) phone recognizers [23] Czech, Hungarian, Russian, German, Hindi, Japanese, and Spanish are used to tokenize the CV segments.

7.1. Sequential Patterns of Phonetic Units

In this analysis, the elements of sequences are the output phonemes decoded by ASR. The PrefixSpan was employed to mine a set of common and unique sequential patterns of phonetic units uttered by the two language groups. For a fair comparison, the output phonemes of all open-set phone recognizers were used to mine the sequential patterns. The minimum support threshold was set to 5 % of the total phonemes decoded by each phone recognizers for each language group. Fig. 8 compares the number of unique sequential patterns of 1-symbol to 5-symbols mined from the two groups. As it can be seen, with increasing length of the sequential patterns, the number of discovered patterns increases exponentially. We can further observe that the *Shang* group generates consistently a higher number of unique patterns compared to the *AE* group for all sequential lengths.

8. Conclusions

This study analyzed child vocalization segments in the newly acquired parallel corpus of sixteen US and Shanghainese toddlers. The corpus captures a realistic audio footprint of a daily life of the subjects. The recordings were manually labeled to avoid segmentation errors due to the frequent interference of the vocalizations with secondary acoustic sources. A set of automatic prosodic and phonotactic analyses was established and used to study various aspects of vocalizations. It was found that the Shanghainese language group tends to, in the majority of vocalizations, produce pitch patterns of longer durations compared to the US group. The interval of the typical first formant occurrence was found broader in the US group. Finally, the Shanghainese group tended to produce a consistently higher number of phone-like patterns. No consistent differences in the preferred formation (contour) of short pitch pattern sequences was observed between the groups. Distribution of speech rhythm in the two language groups was found nearly identical while there was a strong variation among the individuals. The results presented in this paper suggest that the proposed automatic assessment framework is sensitive to various aspects of child vocalization and the authors intend to apply the established methods to analyze a broader group of subjects from the ongoing parallel US/Shanghainese acquisition campaign.

9. 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 *1^{st} 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] 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.

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

[7] 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.

[8] D. Xu, J. Gilkerson, and J. A. Richards, "Objective child vocal development measurement with naturalistic daylong audio recording," in *Proc. Interspeech 2012*, 2012.

[9] L. Polka, S. Rvachew, and K. Mattoc, *Blackwell Handbook of Language Development*. Blackwell Pub., 2006, ch. Experiential Influences on Speech Perception and Speech Production in Infancy. E. Hoff and M. Shatz (Eds.), pp. 153–172.

[10] 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 America*, vol. 72, no. 2, pp. 353–365, 1982.

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

[12] C. Heinrich and F. Schiel, "Estimating speaking rate by means of rhythmicity parameters," in *Proc. Interspeech 2011*, 2011.

[13] K. Ishizuka, R. Mugitani, H. Kato, and S. Amano, "Longitudinal developmental changes in spectral peaks of vowels produced by Japanese infants," *The Journal of the Acoust. Soc. of America*, vol. 121, no. 4, pp. 2272–2282, 2007.

[14] H. R. Gilbert, M. P. Robb, and Y. Chen, "Formant frequency development: 15 to 36 months." *Journal of Voice*, vol. 11, no. 3, pp. 260–266, 1997.

[15] H. Bořil, "Robust speech recognition: Analysis and equalization of Lombard effect in Czech corpora," Ph.D. dissertation, Czech Technical University in Prague, Czech Republic, 2008.

[16] P. A. Torres-Carrasquillo, E. Singer, M. K. R. J. Greene, D. Reynolds, and J. D. Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *Proc. of ICSLP 2002*, 2000, pp. 89–92.

[17] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proc. 1995 Int. Conf. Data Engineering (ICDEf95)*, 1995, pp. 3–14.

[18] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu, "Mining sequential patterns by pattern-growth: the prefixspan approach," *IEEE Trans Knowl Data Eng*, vol. 16, pp. 1424–1440, 2004.

[19] 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.

[20] 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.

[21] 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.

[22] Q. Zhang, H. Bořil, and J. H. L. Hansen, "Supervector preprocessing for PRSVM-based Chinese and Arabic dialect identification," in *Accepted to IEEE ICASSP 2013*, 2013.

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