The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Automatic Detection of Palatalized Consonants in Kashmiri

Ramakrishna Thirumuru[1], *Krishna Gurugubelli*[2], *Anil Kumar Vuppala*[3]

Speech Processing Laboratory, LTRC, KCIS
International Institute of Information Technology, Hyderabad, India
{ramakrishna.thirumuru, krishna.gurugubelli}@research.iiit.ac.in, anil.vuppala@iiit.ac.in

Abstract

In this study, the acoustic-phonetic attributes of palatalization in the Kashmiri speech is investigated. It is a unique phonetic feature of Kashmiri in the Indian context. An automated approach is proposed to detect this unique phonetic feature from the continuous Kashmiri speech. The *i-matra* vowel has the impact of palatalizing the consonant connected to it. Therefore, these consonants investigated in synchronous with vowel regions, which are spotted using the instantaneous energy computed from the envelope-derivative of the speech signal. The resonating characteristics of the vocal-tract system framework that reflect the formant dynamics are used to differentiate palatalized consonants from the other consonants. In this regard, the Hilbert envelope of the numerator of the group-delay function that provides good time-frequency resolution used to extract formants. The palatalization detection experimentation carried out in various vowel contexts using the acoustic cues, and it produced a promising result with a detection accuracy of **92.46 %**.

Index Terms: Kashmiri speech, Palatalized consonants, Envelope derivative, Hilbert envelope of the numerator of the group-delay function

1. Introduction

Kashmiri is referred to as a unique language in the Indian subcontinent that has broad utilization of palatalized sounds in its phonology. Both lexical structures and inflected structures can have them which may appear differently in relation to their non-palatalized partners [1]. Palatalization refers to the process of phone change in which a nonpalatal consonant, changes to a palatal consonant through the movement of the tongue towards hard palate [2, 3]. It occurs as an additional place of articulation of a consonant. This unique phonetic feature detection might be utilized as a part of developing a precise algorithmic approach for pronounciation system in Kashmiri phonology. It can also be utilized to enhance the performance of the language identification system by providing an additional information in terms of unique phonetic features.

The phonemic contrast provided by the palatalization process has been discussed in Russian and other Slavic dialects [4, 5, 6]. In light of an acoustic study of 22 dialects from the Slavic, Celtic, and Uralic families, patterns of cross-linguistic relevance identifying with palatalization is presented in [7, 5]. The palatalization processes are not regular in Indian dialects and Kashmiri emerges in showing an ancillary phenomenon of this feature. All the more particularly, full-palatalization, otherwise called coronalization is much of the time experienced in other Indian languages. The secondary palatalization is reported with instrumental analysis for Kashmiri dialects [1, 8]. It illustrates that the palatalized labials are more set apart than palatalized coronals based on a few distributional examples. In [9], revealed similarities and differences in palatalization patterns due to the place of articulation of target consonants in different language samples. From the literature survey, it is noted that phonological differences of palatalization are realized by various connecting acoustic cues. The articulatory knowledge and acoustic signals to the phonological differentiations are language specific [10].

The objective of this paper is to explore the inborn acoustic cues of palatalization using advanced signal processing techniques that differentiate them from other phones. These acoustic cues are established based on vocal-tract system characteristics during the production of these phones. The vocal-tract can't change it's shape momentarily from the arrangement of one phonetic fragment to that of the following. During the time the articulators are in movement, the resonating characteristics of the vocal-tract are evolving. The change in resonance manifests itself as rising or falling of the formants, or resonance peaks, in the phonic spectrum. An exemplary investigation of natural speech demonstrated that the correct place of articulation of a consonant is influenced by the former and following vocalic condition, with the formant change loci reflecting the changeability of the place of articulation [11]. In this regard, speech formant structure is extracted from the Hilbert envelope of the numerator of the group-delay function (HNGDF). It is derived from a highly decaying windowed speech signal to estimate formants with a better resolution. The acoustic cues for the palatalized consonants are evaluated by anchoring around vowel-regions. This methodology has been used as these phones can occur in vowel-consonant and consonant-vowel contexts.

The remaining paper is arranged in the following manner: Section II describes vowel region detection based on instantaneous energy derived using envelope-derivative of the speech signal. In Section III, a formant tracking algorithm is described utilizing the HNGDF. Section IV presents an automated method to detect palatalized consonants. In Section V, test results to demonstrate the robustness of proposed approach is illustrated. Finishing up comments have been incorporated into Section VI.

2. Vowel region detection

In this work, palatalized consonants are spotted in the continuous speech by anchoring vowel regions. Therefore, an accurate vowel region detection is appreciated in this circumstance. An algorithm for vowel detection can be divided into two phases. The first phase computes the vowel prospect evidence and the second stage estimates the vowel boundaries from the vowel prospect contour.

The critical energy change in excitation source, spectral peaks, and modulation spectrum is seen at the vowel boundaries of the speech signal [12]. Based on these cues, it is hypothesized that a vowel prospect evidence can be obtained from the instantaneous energy contour of the speech signal. In this work,

10.21437/SLTU.2018-25

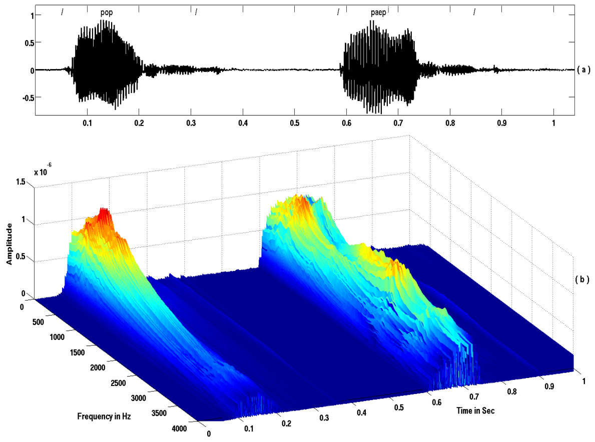

Figure 1: *Contrast in time-frequency representation of the non-palatalized (/pop/) and palatalized (/paep^j/) Kashmiri words using zero-time windowed HNGDF spectrum. (a) Speech signal. (b) Time-frequency representation of speech signal.*

instantaneous energy is estimated using a envelope-derivative of the speech signal. It is a non-negative frequency weighted instantaneous energy, that provides good time-frequency resolution [13]. Considering $s(t)$ as a speech signal with a band limited power spectrum $S_{xx}(\omega) = 0$ for $|\omega| > B$. $s[n]$ is the sampled version of $s(t)$ with sampling duration $T < \pi/B$. The envelope-derivative of the speech signal for the both continuous and discrete time signals is given by using a operator $\varGamma$ [13]. The continuous time nonlinear energy operator is defined as

$$\varGamma\left[s\left(t\right)\right] = \left|\dot{s}\left(t\right) + jH\left[\dot{s}\left(t\right)\right]\right|^2 = \dot{s}^2\left(t\right) + H\left[\dot{s}\left(t\right)\right]^2 \quad (1)$$

where $\dot{s}\left(t\right)$ is the first-order derivative of $s(t)$ and $H\left[.\right]$ correspond to the Hilbert transform operator. For the discrete time case $\varGamma_d$ is defined as

$$\begin{aligned}\varGamma_d[s[n]] =& \frac{1}{4}[s^2[n+1] + s^2[n-1] + h^2[n+1] + h^2[n-1]] \\ &+ \frac{1}{2}[s[n+1]s[n-1] + h[n+1]h[n-1]] \quad (2)\end{aligned}$$

where $h\left[n\right]$ is the Hilbert transform of the signal $s\left[n\right]$. This operator facilitates to compute instantaneous energy using three samples. Thus obtained evidence is subjected to the enhancement using first order difference of the same. In the next phase, the enhanced energy profile is convolved with a first-order Gaussian differentiator and the peaks and valleys of the convolved output are hypothesized as the vowel start and end points respectively. Lastly, the excitation source information like uniformity of epoch intervals and the strength of the excitation [14, 15] are computed from the zero frequency filtered signal [15, 16, 17]. These additional cues are utilized in removing spurious regions and correcting the hypothesized vowel boundaries [18].

3. Formant extraction using the HNGDF

Traditional spectral analysis for the short segments of speech experience poor frequency resolution due to the limitation of window sizing. Thus, high-resolution characteristic of Hilbert envelope of the numerator group-delay based technique can be used for extracting spectral content from short segments of speech [19]. The basic procedural steps involved in this method are depicted as follows:

1. A short segment of size M is selected from the pre-emphasized speech signal denoted by $s[n]$.
2. Discrete Fourier transform of length N greater than M is chosen by adding the required number of zeros to the speech segment.
3. A zero-time windowed signal is computed and denoted by $x[n] = s[n].w[n]$, where $n = 0, 1, 2, ..., N-1$ and zero-time window is given by

   $$w[n] = \begin{cases} 0, & n = 0 \\ \frac{1}{4sin^2(\pi n/(2N))}, & n = 1, 2, .., N-1. \end{cases} \quad (3)$$

   The truncation effect due to zero-time window is compensated using another tapering window given by $w_1\left[n\right] = 4cos^2\left(\frac{\pi n}{2N}\right), n = 0, 1, 2, ..., N-1$, and N is the length of the window.
4. The numerator of the group-delay function is estimated for windowed signal [20] and it is represented as

   $$g\left[k\right] = X_R\left[k\right]Y_R\left[k\right] + X_I\left[k\right]Y_I\left[k\right], k = 0, 1, .., N-1 \quad (4)$$

   $X\left[k\right] = X_R\left[k\right] + jX_I\left[k\right]$ is the discrete Fourier transform of $x\left[n\right]$, and $Y\left[k\right] = Y_R\left[k\right] + jY_I\left[k\right]$ is the discrete Fourier transform of $y\left[n\right] = nx\left[n\right]$. The zero-time windowed numerator of the group-delay function brings out the spectral content superior to the conventional magnitude spectrum.
5. By progressively differencing the numerator of the group-delay function spectral content of the speech is highlighted. The crests in the differenced numerator of the group-delay spectrum are more likely to be affected by the nearby spectral troughs due to large bandwidths. In order to highlight the peaks of the spectral features, it is further processed by computing the Hilbert envelope of the differenced numerator of the group-delay spectrum. Thus spectral peaks of short speech segment are extracted using HNGDF.

The spectrogram for a Kashmiri utterance having palatalization and non-palatalization is shown in the Figure. 1. Two conclusions can be drawn from this illustration. A spectral change is perceived around 1500 Hz and a reasonably high frequency energy is observed above 2000 Hz in the utterance having palatalized consonant. These remarks are in-line with the studies carried out on Kashmiri palatalization [21, 1] and a signal processing measure can be formulated to detect these phones in a continuous speech.

4. Proposed approach for the detection of palatalized consonant

The acoustic cues for detecting the palatalized consonant in continuous speech are discussed in this section. According to the Kashmiri linguistic theory, palatalization can occur in the *i-matra* vowel context. It is by and large concurred that the most illustrative parameter separating amongst palatalized and non-palatalized consonants is the locus of the second formant (F2) in consonant-vowel or vowel-consonant transition [21]. The F2 rise in the connected vowel region leading to the palatal region

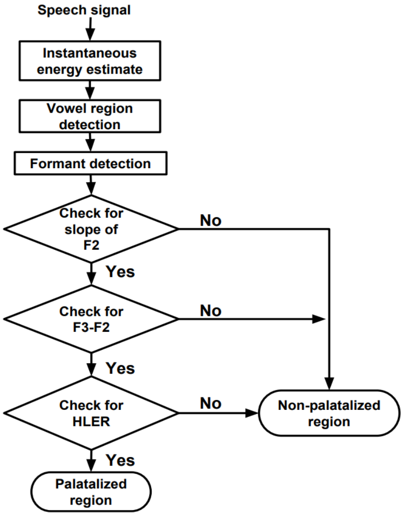

Figure 2: *Flow diagram.*

is clearly noted from the HNGD spectrum. This characteristic may be attributed to the tongue-rise during palatalization. The offset portion of Kashmiri palatalization is typically a voiceless front unrounded approximant that includes vowels such as [i], [u] and a semi-vowels [j] [1, 22]. A turbulent air stream is produced during their production and it is observed from the HNGDF spectrum as a fricative noise above 2 kHz. Therefore, the vowel regions are extracted using the algorithm discussed in Section II. In the next level, the regions on the either side of the vowel regions are processed to spot palatalized consonants in the continuous speech. The acoustic cues used in this study are gradient of F2 in the vowel region connected to the phone (G), frequency difference between F3 and F2 (D), and high to low energy ratio ($HLER$) on either side of vowel region. The averaged gradient (G) pertaining to the F2 is estimated as

$$G = \frac{F_{ep} - F_{op}}{t_{ep} - t_{op}} \tag{5}$$

where F_{ep} is the formant magnitude at the vowel end-point, F_{op} is the formant magnitude at the vowel onset point, t_{ep} time at the vowel end-point and t_{op} is the time at vowel onset point. The high to low energy ratio ($HLER$) is given by

$$HLER = \frac{\sum_{k=2001}^{4000} h(k)}{\sum_{k=0}^{2000} h(k)} \tag{6}$$

where $h(k)$ denotes spectral energy at k^{th} frequency. Alongside these cues, frequency contrast measure (D) of second and third formants (F3-F2) in the selected region is given by

$$D = min(||F3 - F2||) \tag{7}$$

The combined measure is thought about with a proper thresholding to group palatalized consonants and non-palatalized consonants. The algorithmic flow chart is depicted in the Figure. 2. A test case is shown in Figure. 3. This figure illustrate the detection mechanism of palatalized consonants. Figure. 3a depict the speech signal along with the ground truths marked by phonetician. Figures. 3b illustrate hypothesized vowel regions in the continuous speech. Figure. 3c and Figure. 3d correspond to the formant structure and predicted palatalized consonants respectively.

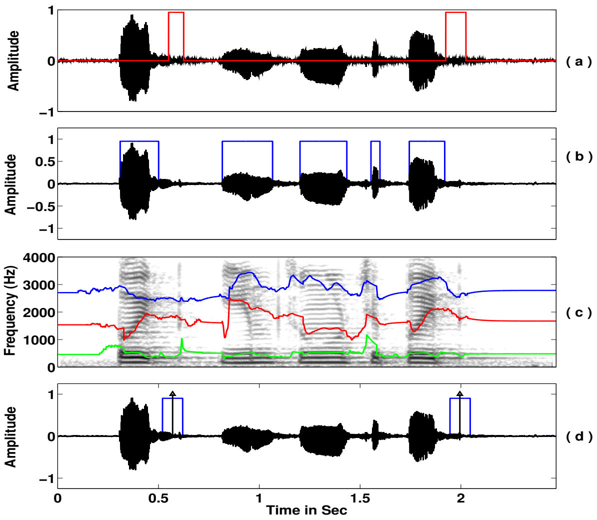

Figure 3: *Detection of palatalized consonants for a Kashmiri speech utterance/'Yem chount chi paep^j'/ (a) Speech signal with ground truths. (b) Predicted vowel regions. (c) Formant tracking using HNGDF and spectrogram in the background. (d) Predicted palatalized consonants.*

5. Experimentation and Results

The palatalized consonant detection algorithm implemented using MATLAB software on a customized dataset. Ten native speakers comprising five men and five women of Kashmir aged between 20 and 30 participated in this experimentation. All speakers are students at the University of Hyderabad, Hyderabad, India. All the members were born and brought up in Kashmir, India. The articulations for the present work comprises of 17 meaningful Kashmiri sentences involving palatalized consonants. All voice recordings were digitized with a higher sampling rate of 44.1 kHz for better sound quality and these recordings are down-sampled to 8 kHz. A few Kashmiri sentences used in the experimentation are listed below.

- *Yem count chi pap (These apples are ripe).*
- *Bi khami zi beak batti (I will eat two handfuls of food).*
- *Mae zayi nichi (I gave birth to a baby girl).*
- *Yetath chu sup (Here is a winnowing basket).*
- *Che chukh bouch (You are a glutton).*

As a part of this experimentation, precise vowel regions are detected using the algorithm discussed in the Section II. In perspective of articulatory to acoustic signal mapping of palatalization in Kashmiri discourse, F2 gradient (G), high to low energy($HLER$), and frequency difference measure (D) on either side of a vowel is computed to discriminate palatalization associated with the consonant connected to the vowel. The combinations of these parameters were plotted in the Figure. 4 to

Table 1: *Detection rate (in %) of palatalized consonants using different second formant gradient measures.*

F2 gradient(G)	Detected	Missed	False Alarm
1.0	96.24	3.76	18.25
0.95	96.24	3.76	18.25
0.90	92.31	7.69	16.54
0.85	90.27	9.63	15.88

study the feasibility of these cues for detecting the palatalization. The cues precedence used for the detection mechanism has been decided upon the relative strength by which they correlate to the underlying phonological process. With various arrangement of threshold limits for the F2 gradient, continuous speech utterances were tested on the framework for spotting palatalized consonants. The recognition technique contrasts areas of palatalized consonants and the ground realities, which are physically marked by phonetician for verification. A set of few outcomes for this approach are accounted for in Table 1. It is observed that a decent recognition rate (96.24%) is accomplished at G = 1.0 and 0.9. In any case, a false alert rate over 18% is noted. This approach is improved by considering frequency difference measure (D) as a second evidence. The speech utterances were tested for different limiting values of frequency difference measures fixing G = 1.0. The results pertaining to this experimentation are illustrated in Table 2. It was observed that better detection rate (94.70%) is achieved with reduced false alarm rate (11.94%) for D = 150. On combined experimentation by fixing G = 1.0 & D=150 Hz, the performance of this framework has been further enhanced with third cue named as $HLER$. This relates to the ratio of high frequency energy above 2 kHz and low frequency energy. The results are shown in the Table 3. These results convey that this system produced optimal detection rate of **92.46%** with a false alarm rate of **7.71%**. This approach detects one of the language-specific phonetic features of Kashmiri with speaker variability and can upgrade the performance of language identification system in Indian context using a set of signal processing measures.

Table 2: *Detection rate (in %) of palatalized consonants using different frequency difference measures with G = 1.0.*

min(F3-F2)(D) in Hz	Detected	Missed	False Alarm
10	90.19	9.81	10.73
100	92.15	7.85	12.36
150	94.70	5.30	11.94
200	94.12	5.88	11.07

Table 3: *Combined detection rate (in %) of palatalized consonants using different high to low energy ratio measures with G = 1.0 & D=150 Hz.*

$HLER$	Detected	Missed	False Alarm
0.1	94.70	5.30	10.7
0.15	94.28	5.72	9.38
0.2	**92.46**	**7.54**	**7.71**
0.25	88.35	11.65	6.94

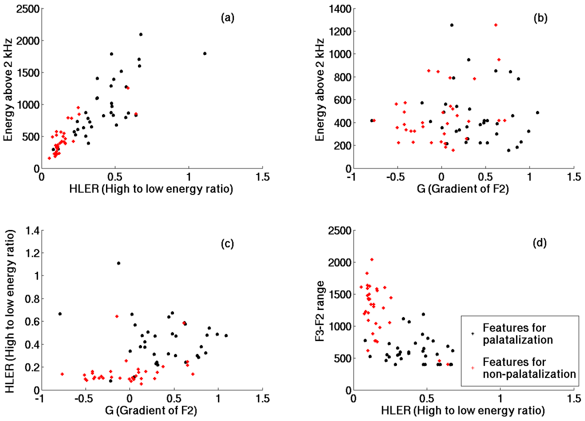

Figure 4: *Feature analysis (a) $HLER$ Vs Energy above 2 kHz plot. (b) Gradient of F2(G) Vs Energy above 2 kHz plot. (c) Gradient of F2(G) vs $HLER$ plot. (d) $HLER$ vs Frequency difference measure$D$ plot.*

6. Summary and conclusions

In this paper, the characteristics of palatalized consonants in Kashmiri was studied. According to the Kashmiri phonology, *i-matra* vowel palatalizes the consonant connected to it. Therefore, the vowel regions have been detected using envelope-derivative of the speech signal that provides a non-negative frequency weighted instantaneous energy contour with good temporal and spectral resolution. The dynamic spectral attributes of these consonants were contemplated utilizing the zero-time liftering technique for examination of Kashmiri speech signal. Through this technique, the spontaneous response of the vocal-tract system has been acquired. This strategy for examination empowered to inspect the points of interest of the phantom highlights (F2 gradient, A measure of F2 nearing F3, and high to low energy ratio) during the palatalization process in the vowel context. These measures were derived from the spectral characteristics of the palatalized consonants utilizing the Hilbert envelope of the numerator of the group delay function. The results showed that acoustic cues such as F2 slope, F3-F2 difference, and high-low energy ratio capture secondary articulatory changes in palatalized consonants to detect them in continuous speech.

7. Acknowledgement

The authors thank Prof. Peri Bhaskararao for providing insights of palatalized consonants in Kashmiri language.

8. References

[1] P. Bhaskararao, S. Hassan, I. A. Naikoo, P. A. Ganai, N. H. Wani, and T. Ahmad, "A phonetic study of kashmiri palatalization," *Working Papers in Corpus-based Linguistics and Language Education*, vol. 3, pp. 1–17, 2009.

[2] D. N. Bhat, "A general study of palatalization," *Universals of human language*, vol. 2, pp. 47–92, 1978.

[3] T. Bynon, *Historical linguistics*. Cambridge University Press, 1977.

[4] G. Fant, *Acoustic theory of speech production: with calculations*

based on X-ray studies of Russian articulations. Walter de Gruyter, 1971, vol. 2.

[5] A. Kochetov, “Production, perception, and emergent phonotactic patterns: A case of contrastive palatalization (2001),” *Toronto Working Papers in Linguistics*, 2001.

[6] B. Connell, “Ladefoged peter. elements of acoustic phonetics, chicago: University of chicago press. 1996. pp. viii+ 216. us 39.95$(hardcover)$,14.95 (softcover).” *Canadian Journal of Linguistics/Revue canadienne de linguistique*, vol. 44, no. 1, pp. 53–55, 1999.

[7] A. Kochetov, “Phonotactic constraints on the distribution of palatalized consonants,” *Toronto Working Papers in Linguistics*, vol. 17, 1999.

[8] P. Bhaskararao, “Salient phonetic features of indian languages in speech technology,” *Sadhana*, vol. 36, no. 5, pp. 587–599, 2011.

[9] N. Bateman, “A crosslinguistic investigation of palatalization,” Ph.D. dissertation, Univ. of California San Diego, 2007.

[10] M. Ordin, “Palatalization and intrinsic prosodic vowel features in russian,” *Language and speech*, vol. 54, no. 4, pp. 547–568, 2011.

[11] S. E. Öhman, “Coarticulation in vcv utterances: Spectrographic measurements,” *The Journal of the Acoustical Society of America*, vol. 39, no. 1, pp. 151–168, 1966.

[12] S. M. Prasanna, B. S. Reddy, and P. Krishnamoorthy, “Vowel onset point detection using source, spectral peaks, and modulation spectrum energies,” *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 556–565, 2009.

[13] J. M. O’Toole, A. Temko, and N. Stevenson, “Assessing instantaneous energy in the EEG: A non-negative, frequency-weighted energy operator,” in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE.* IEEE, 2014, pp. 3288–3291.

[14] P. Gangamohan, S. R. Kadiri, S. V. Gangashetty, and B. Yegnanarayana, “Excitation source features for discrimination of anger and happy emotions,” in *Proc. Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[15] B. Yegnanarayana and K. S. R. Murty, “Event-based instantaneous fundamental frequency estimation from speech signals,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 614–624, 2009.

[16] K. S. R. Murty and B. Yegnanarayana, “Epoch extraction from speech signals,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[17] B. Yegnanarayana, S. M. Prasanna, and S. Guruprasad, “Study of robustness of zero frequency resonator method for extraction of fundamental frequency,” in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.* IEEE, 2011, pp. 5392–5395.

[18] R. Thirumuru, S. V. Gangashetty, and A. K. Vuppala, “Improved vowel region detection from a continuous speech using post processing of vowel onset points and vowel end-points,” *Multimedia Tools and Applications*, pp. 1–15, 2017.

[19] Y. Bayya and D. N. Gowda, “Spectro-temporal analysis of speech signals using zero-time windowing and group delay function,” *Speech Communication*, vol. 55, no. 6, pp. 782–795, 2013.

[20] J. M. Anand, S. Guruprasad, and B. Yegnanarayana, “Extracting formants from short segments of speech using group delay functions,” in *Proc. Ninth International Conference on Spoken Language Processing*, 2006.

[21] R. AMBRAZEVIČIUS, “Differential acoustical cues for palatalized vs nonpalatalized prevocalic sonants in lithuanian.” *Man & the Word/Zmogus ir zodis*, vol. 12, no. 1, 2010.

[22] J. J. McCarthy, *The phonetics and phonology of Semitic pharyngeals.* PA, 1994.