An acoustic-articulatory database of VCV sequences and words in Toda at different speaking rates

Shankar Narayanan  Aravind Illa  Nayan Anand  Ganesh Sinisetty Karthick Narayanan Prasanta Kumar Ghosh
IISc Bangalore  IISc Bangalore ICFAI Hyderabad  IISc Bangalore  JNU Delhi  IISc Bangalore

Abstract—We present a database comprising simultaneous acoustic and articulatory recordings of thirty $V_1CV_2$ nonsense words and forty two Toda words, recorded with an electromagnetic articulograph and spoken by six Toda speakers (two males and four females) at four different speaking rates, namely slow, normal, fast and very fast. The vowels in the $V_1CV_2$ ($V_1 \neq V_2$) come from a set of six vowels, namely, /a/, /l/, /l/, /l/, /l/, /l/, /l/, where the last vowel is a front rounded vowel in Toda. The consonant in the $V_1CV_2$ stimuli is chosen as /p/ for all the recordings. The articulatory data in the proposed database comprises recording of movements of five articulatory points, namely, upper lip, lower lip, jaw, tongue tip and tongue dorsum in the midsagittal plane. The acoustic and articulatory recordings are made available at 16 kHz and 100 Hz respectively. Boundaries of vowels and consonant in $V_1CV_2$ stimuli are provided along with this database. Basic acoustic and articulatory analysis of the $V_1CV_2$ recordings in this database are presented, which show the manner in which the acoustic and articulatory spaces as well as coarticulation change with speaking rates. The proposed database is suited for a number of research studies including the effect of speaking rates on the acoustic and articulatory aspects of coarticulation in Toda, analysis of labial kinematics during consonant production at different speaking rates, and acoustic-articulatory analysis of front rounded vowel in Toda.

Index Terms: electromagnetic articulograph, speaking rate, coarticulation

1. INTRODUCTION

Speech is produced as a result of phoneme sequence being spoken continuously. Thus, unlike isolated phonemes, the presence of the overlay effect of one phoneme over another causes coarticulation, blurring their intermediate boundaries [1]. The coarticulation is manifested both in the acoustic and articulatory domains. In acoustic domain, the spectro-temporal characteristic of each phoneme is influenced by its neighbouring phonemes. On the other hand, the articulatory gestures [2] for each phoneme overlap with that of its neighbouring phonemes. In running speech, the acoustic and articulatory targets for each phoneme are achieved at varying precision and degree depending on the degree, magnitude and directionality of the coarticulation. These depend on a number of factors including the degree of frontness or backness of the distribution of vowel spaces [3], the height of the vowels [4], vowel specific consonantal effect [5] and speaking rate. Coarticulation has an important role to play in several studies and applications including lip reading [6], the synthesis of visual speech [7] or word recognition for toddlers. [8]

Speaking rate not only varies across speakers but also depends on the context, content and emotional state of a speaker. As speaking rate is altered, it has a complex effect on the acoustic and articulatory properties of the phonemes and coarticulation among them. It is known that the vowel space changes with speaking rate. For example, Gay T [9] reported changes in the vowel formants with speaking rates. It is also known that the variability in target articulatory movement for each phoneme reduces with increasing speaking rate [10]. As a result of the change in the acoustic and articulatory characteristics of the phonemes, the degree and nature of coarticulation also changes at different speaking rates [11].

Coarticulation among vowels and consonants has typically been studied with nonsense words, i.e., VCV sequences with varying number of subjects. The gender, age, and native languages of the subjects considered also vary depending on the study. Table I summarizes the VCV stimuli and subject details from a set of works in the literature which use nonsense VCV sequence for different studies. It is clear that a nonsense VCV word is a common choice for study on coarticulation.

A few coarticulation studies have also been done with recordings of meaningful words and sentences. For example, second formant was used to analyze the coarticulation in CV syllables occurring in the spontaneous speech of five speakers and in the same words read in isolation [21]. Thus, it is evident that coarticulation studies have been primarily confined to nonsense words recorded from speakers from a number of languages including English, Russian, Swedish, Italian, and Arabic. Coarticulation with language-specific meaningful words could bring out coarticulation characteristics specific to the language. Also, very few works in the literature focus on understanding of the effect of rate on coarticulation. Similarly, coarticulation studies have rarely been done based on directly
measured articulatory movements. However, no work exists where the effect of speaking rate has been studied on the acoustic and articulatory aspects of coarticulation.

In this work, we focus on collecting a database which addresses both acoustic and articulatory recordings for an under-documented language called Toda. Toda is a native tribe of Nilgiri hills having a sparse population of around 1200 members. Furthermore Toda has been experiencing an outbursting dissimilation and vowel shift [22] leading to loss of unique distinctive Toda speech sounds. The dataset, proposed in this work, contains acoustic-articulatory recordings using electromagnetic articulograph (EMA) from six Toda speakers (two males and four females) speaking thirty \( V_1pV_2 (V_1 \neq V_2) \) stimuli where \( V_1, V_2 \in \{/a/, /e/, /i/, /o/, /u/, /y/\} \) at different speaking rates. Apart from the nonsense words, we have also recorded forty two Toda words, characterized by unique Toda sounds as well as speaking style of Toda speakers.

The database proposed in this work is suited for a number of potential research studies including the effect of speaking rates and speaking styles (nonsense and meaningful words) on the acoustic and articulatory aspects of coarticulation, analysis of labial kinematics during consonant production at different speaking rates, investigation of the speaker-specific and speaker-independent factors in coarticulation. Acoustic and articulatory analysis of Toda front rounded vowel can be carried out with this database. As articulatory motion recording at different speaking rates is a unique feature of this database, it also enables studies on speaking rate influence on voice onset time as well as consonantal effect over vowel formants, oral closure pattern as a function of speaking rate.

II. DATABASE COLLECTION

A. Toda speakers and speech stimuli

The database contains acoustic and articulatory recordings from six Toda speakers comprising two male (TM1, TM2) and four female (TF1, TF2, TF3, TF4) speakers aged 49, 28, 61, 24, 16 and 23 years respectively. Speakers TM1, TM2, TF1, TF2 and TF3 belong to “towoolhythoolh” and speaker TF4 belongs to “tortas” tribal sub-community of Nilgiris district in Tamil Nadu, India. All the speakers were native speakers of Toda and Tamil (regional language of the state) and reported to have no speech disorders in the past. All subjects signed the consent form as recommended by the institute ethics committee. All recordings were collected at SPIRE lab speech recording room facility in Electrical Engineering, Indian Institute of Science, Bangalore, India.

The acoustic and articulatory data are collected using nonsense and real world stimuli. The nonsense stimuli comprise 30 consonant-vowel-consonant-vowel (CV1CV2) asymmetrical \( (V_1 \neq V_2) \) sequences and 42 Toda words. An initial consonant was added to the VCV sequences to avoid coarticulation between \( V_1 \) and the vowel preceding it in the carrier phrase. The VCV sequences were constructed using six vowels and one consonant. The six vowels consist of one close front rounded Toda vowel /y/ [23], two front vowels /i/, /e/, one mid vowel /a/ and two back vowels /u/, /o/. One labial consonant /p/ was considered. In \( C_1V_1C_2V_2 \) stimulus, \( C_1 \) and \( C_2 \) were chosen to be /p/, and \( V_1 \) was varied over the six vowels. \( V_2 \) was varied over the six vowels excluding the vowel used in \( V_1 \). This results in a set of \( 30 \times 6 \times 1 \times 5 \) asymmetrical \( pV_1pV_2 \) sequences. Speakers were asked to speak each \( pV_1pV_2 \) stimulus in a carrier phrase as follows: “own/ \( pV_1pV_2 \) hidispinfi” (“I say \( pV_1pV_2 \)”). All the 30 \( pV_1pV_2 \) stimuli were recorded at 4 different speaking rates (slow, normal, fast and very fast) with 3 repetitions for each. Overall, 2160 (6 speakers \( \times \) 30 VCV combinations \( \times \) 4 speech rates \( \times \) 3 repetitions) \( pV_1pV_2 \) stimuli were recorded. For real world stimuli we consider 42 words in Toda language. Table II lists the set of words that were recorded and the corresponding English meaning and phonetic transcriptions.

<table>
<thead>
<tr>
<th>Toda word (English)</th>
<th>IPA</th>
<th>Toda word (English)</th>
<th>IPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>APORSHI (tumid)</td>
<td>/pa ah th ih/</td>
<td>VUDI (only)</td>
<td>/v oy dh ih/</td>
</tr>
<tr>
<td>AMEDHII (peace)</td>
<td>/ma ed ah th ih/</td>
<td>BADIL (answer)</td>
<td>/b ah ah th l/</td>
</tr>
<tr>
<td>AKOH (her)</td>
<td>/ak oh hh/</td>
<td>ADEMODIRI (same)</td>
<td>/nd ah m oh dh ih r ih/</td>
</tr>
<tr>
<td>AGRUS (accusing)</td>
<td>/a g ah s/</td>
<td>AROD (sassy)</td>
<td>/a r oih dh/</td>
</tr>
<tr>
<td>VIVARCHH (describe)</td>
<td>/v ih v ah ah th/</td>
<td>ADHURH (because)</td>
<td>/ah dh ih v dh/</td>
</tr>
<tr>
<td>REVHCHE (require)</td>
<td>/r eh v eh y ah/</td>
<td>VIVFURM (default)</td>
<td>/v ah v ah r m/</td>
</tr>
<tr>
<td>INNURASHI (recently)</td>
<td>/in n uih oh m/</td>
<td>SERI (ok)</td>
<td>/s ah r ih/</td>
</tr>
<tr>
<td>NIMUSHIM (minute)</td>
<td>/n uih th uih ah ah th/</td>
<td>EDVASHL (second)</td>
<td>/ae ah v ah ah th/</td>
</tr>
<tr>
<td>REVE (need)</td>
<td>/r e vi th y ah th/</td>
<td>METONI (step)</td>
<td>/m ah t oh l n th/</td>
</tr>
<tr>
<td>NEPAUTHI (recognize)</td>
<td>/n e p a ah s th/</td>
<td>PEERU (reach)</td>
<td>/p a e uh/</td>
</tr>
<tr>
<td>ENODI (tool)</td>
<td>/e no uih y ah th/</td>
<td>PORDIP (effect)</td>
<td>/p oh dh ih p/</td>
</tr>
<tr>
<td>UDURUITH (against)</td>
<td>/ah dih r th ih th/</td>
<td>SERTHI (store)</td>
<td>/s oih y ah th/</td>
</tr>
<tr>
<td>MEDIUVASHI (determine)</td>
<td>/ih th uih th y ah ah th/</td>
<td>SOMALETS (manage)</td>
<td>/s oh ma l i s/</td>
</tr>
<tr>
<td>KOPERSH (sparrow)</td>
<td>/k oh p ah th/</td>
<td>OPUM (increase)</td>
<td>/o h p uh m/</td>
</tr>
<tr>
<td>KOPOITH (present)</td>
<td>/k oh p ah th ih th/</td>
<td>PUS (bad)</td>
<td>/p ah y ih/</td>
</tr>
<tr>
<td>POUDI (common)</td>
<td>/p oh dh uih th/</td>
<td>MUSCLE (nice)</td>
<td>/m u sh ah ah/</td>
</tr>
<tr>
<td>UDITHI (help)</td>
<td>/uh dih ih th/</td>
<td>ANUBHAVAM (experience)</td>
<td>/n uih ah th ah v ah m/</td>
</tr>
<tr>
<td>KUTER (Sacred tree)</td>
<td>/k uh t ah r/</td>
<td>HORTHUU (trouble)</td>
<td>/uh b oh uih uih r uh/</td>
</tr>
<tr>
<td>MUELEH (elf)</td>
<td>/m uih y ah ah th/</td>
<td>MURCEUY (fly)</td>
<td>/m uih y c ah y th/</td>
</tr>
<tr>
<td>UDINOUM (although)</td>
<td>/uh th y oh y uih m/</td>
<td>INOH (what)</td>
<td>/ih n oh bh/</td>
</tr>
<tr>
<td>VECHOUTHI (successful)</td>
<td>/v eh th oih y ih th/</td>
<td>ARIZINGUD (yellow)</td>
<td>/ar z ih ng uih dh/</td>
</tr>
</tbody>
</table>

B. Recording setup and protocol

For each stimulus, simultaneous audio and articulatory movement data were recorded. To record articulatory movements, EMA AG501 [24] was used with five sensors placed on different articulators, namely Upper Lip (UL), Lower Lip (LL), Jaw, Tongue Tip (TT) and Tongue Dorsum (TD). Two additional sensors were placed behind the two ears for correcting head motion [25]. An illustrative diagram of the sensor placement is shown in Fig. 1. Each sensor captures the motion of articulators in 3D spatial coordinates with a precision of 0.3mm (RMS) and with a sampling rate of 250Hz [24]. A t.bose EM9600 shotgun, unidirectional electret condenser microphone EM9 [26] was placed at a distance of 1-2 feet from the subject to record the audio data with a sampling rate of 48 kHz synchronously with the articulatory data. The speakers were asked to seat comfortably in the EMA recording setup. Then, the sensors coated with “plasty-late” latex material [27] was glued using “Epiglu” to the articulators.
of the speaker. The speakers were given sufficient time to get accustomed to speaking with sensors attached and produce natural speech.

Speakers were instructed with the recording protocol before beginning of recording and also during the course of recording as described below. Speakers were instructed to read every stimulus aloud at self-determined speaking rates, beginning with a clear and slow articulation labeled as “Slow”, and then with a habitual rate, labeled “Normal”; and then with a faster rate labeled “Fast”; followed by the fourth rate condition labeled as “Vfast” in which speakers were instructed to speak as fast as possible without sacrificing intelligibility or clarity. Four levels of speaking rates were used to ensure maximum speaking rate discrimination between the first and fourth levels. Each stimulus was repeated three times by every speaker in each of the four speaking rates in order of slow, normal, fast and very fast in a single stretch, which allows each speaker to judge and adjust their speaking rate in the required order. A single recording session took approximately two hours on an average. A Graphical User Interface (GUI) was designed for this purpose to provide the speakers with visual and audio cues to vary speaking rates. Each stimulus was spoken 12 times (4 rates × 3 repetitions per rate) by the speakers. A beep, as an audio cue, was played every time before a speaker repeats a stimulus. The beep sound was fixed over 3 repetitions within a rate and it varies across the four rates. A rising chirp signal was used as the beep sound which starts at a frequency of 500Hz and ends at 2000Hz, with varying duration across different rates, with the longest duration for Slow and the shortest for Vfast rate. A visual cue was also displayed on GUI, with a horizontally moving bar where the bar width was varied simultaneously with the played beep sound.

C. Post processing and data annotation

To reduce computational complexity, the recorded audio was downsampled to 16 kHz and articulatory data was downsampled to 100 Hz. Further the recorded articulatory data was low-pass filtered with a cut-off frequency of 40 Hz [10] to avoid high frequency noise resulting from EMA measurement error. Vowel consonant boundaries were manually marked for each pV1pV2 stimulus to remove the carrier phrase. Vowel offset measurements for V1 were made from visual examination of the wideband spectrogram and speech signal to determine the last glottal pulse of the vowel; V2 onset was taken at the first glottal pulse following the release burst of the consonant. All onsets and offsets were assessed with a magnified view of the audio/spectrographic signal. After extracting pV1pV2 stimuli, F1 and F2 were computed at a rate of 500Hz with 20ms window length and 2ms shift using Praat [29] based on linear predictive coding peak-picking method [30]. For the words stimuli, we perform manual annotation to remove begin and end silences. Fig. 2, illustrates the spectrogram and articulatory trajectories of TTZ and Lip Aperture (LA) (distance between UL and LL) of a sample Toda word ‘ADEMODIRI’. From Fig. 2, at phoneme /m/, LA approaches zero due to lip closure. For /d/ and /t/, due to tongue tip constriction, TTZ contour shows local maxima. We observed that articulatory trajectories corresponding to Vfast speaking rate tend to have less variations compared to that of Slow speaking rate.

III. PRELIMINARY ANALYSIS ON DATABASE

Duration analysis of pV1pV2 across speaking rates: We computed the total duration pV1pV2 and individual duration of vowels at V1 and V2 region at different speaking rates. Fig. 3 illustrates the distribution of pV1pV2 duration across speaking rates using a box plot. We compute the average duration of all six vowels across all speakers and repetitions in both V1 and V2 and at both rates (Slow and Vfast) separately. Table III reports the average of these duration values along with standard deviation (SD) in bracket. We observe that, from Slow to Vfast, there is a consistent drop in duration for both V1 and V2. t-test is conducted to measure the difference in means of the duration in Slow vs Vfast speaking rates. Significant (p < 0.05) drop in the vowel duration is observed from Slow to Vfast rate for both V1 and V2 for all subjects in majority of the cases.
change in speaking rate from Slow and Vfast rates. The arrow in each figure indicates the change in the formant frequencies for a vowel when speaking rate is changed from Slow (circle) to Vfast (star). Further, we observe that, from Slow to Vfast speaking rate, there is a reduction in vowel space.

Vowel space: The vowel space represented by F1-F2 coordinates is known to vary with speaking rates. With an increase in speaking rate Augustine Agwuele et al. [31], have shown that there is a reduction in vowel space. We present the analysis on the impact of speaking rates on V1 in VpV2 stimuli with acoustic and articulatory data. Fig. 4 illustrates the scatter plot of TT and TD in X (x-axis) and Z (y-axis) direction for phone /a/ and /o/ in V1 region for all the subjects, where elliptical contours indicate the co-variance of the respective phone assuming a normal probability distribution. From Fig. 4 (b), we observe that there is a shift in distribution towards the back side from Vfast to Slow speaking rate. Since phone /o/ is a back vowel and at fast speaking rate there is a drop in the extent of gestural movements. This could be due to the fact that an increase in rate is achieved by increasing the velocity and/or decreasing the range of articulatory movement [32]. This results in higher overlap among articulatory gestures [33]–[35] and vowel space reduction [36]. To observe the rate effect in acoustics, the median values of F1 and F2 in the V1 nucleus are used for obtaining the vowel space. Fig. 5 illustrates change in vowel space for all speakers with the

Authorized licensed use limited to: J.R.D. Tata Memorial Library Indian Institute of Science Bengaluru. Downloaded on December 31,2020 at 11:12:25 UTC from IEEE Xplore. Restrictions apply.
IV. POTENTIAL RESEARCH AND DEVELOPMENT USE

As the database reported in this work contains both $V_1pV_2$ and words spoken at different speaking rates, it opens up opportunities to address a number of research questions using simultaneous acoustic and articulatory data. A few of these potential research and development uses are summarized in the following subsections.

A. Effect of speaking rate on coarticulation

Understanding the impact of speaking rate on the coarticulation has been a research question of interest for several decades [38]–[40]. Although it is generally known that a slow speaking rate results in decrease in coarticulation, how the speaking rate impacts the anticipatory and carryover coarticulation in different $V_1-V_2$ context is not clear. While most of the existing studies use only acoustic representation using locus equation, a joint modeling of acoustic-articulatory variable would uncover the dynamics of the speech production during co-articulation in a principled way. The database presented in this paper would be an ideal resource for such a study.

B. Effect of speaking style on coarticulation

While most of the studies on coarticulation have been done using controlled VCV sequences [12], [15], few works have examined how coarticulation characteristics in other speaking style vary. For example, a study by Judit et al. [41] examined voicing assimilation in different speaking styles in Hungarian: reading, interpreted and spontaneous speech. Unlike VCV, coarticulation studies have also focused on words. In particular, Maolin Wang [42] examined trans-segmental vowel-to-vowel anticipatory coarticulatory effect in Chinese. The database presented in this paper provides a unique set of recordings which can be used to study CV and VC (with C=/p/) coarticulation both in isolated $V_1CV_2$ sequence as well as in Toda words. While $V_1CV_2$ sequence is an artificial nonsense stimulus, the words, being part of the everyday use in Toda language, may illustrate different aspects in coarticulation for the same VC and CV. It would be worth investigating whether the findings on coarticulation from $V_1CV_2$ sequence hold true for words as well.

C. Study of labial kinematics during consonant production at different speaking rates

Understanding the dynamics of lips during production of labial stop consonant in nonsense VCV sequence is of great interest in many investigations of speech motor control [43]–[45]. In order to carry out studies on lip kinematics during production of /p/ at a large number of vowel context in $V_1pV_2$ sequence, the database in this paper would be ideal as it provides recordings of $V_1pV_2$ sequence from six subjects in 30 $V_1-V_2$ combinations. This database is also unique when such a study on labial kinematics has to be carried out to compare the characteristics at different speaking rates. With this database, one can also compare lip kinematics in $V_1CV_2$ sequence and isolated Toda words.

D. An acoustic-articulatory study on speaker specific and speaker independent factors in coarticulation

As the database contains $V_1pV_2$ sequences and Toda words from six subjects, it would be a rich source for studying aspects in coarticulation that occur independent of the speaker and also appear in a speaker-specific manner. The acoustic-articulatory modalities in this database would be useful to carry out such study in both acoustic and articulatory domain. Speaker-specific characteristics in coarticulation have been reported in the past [46]. In fact, nasal coarticulation has been used for speaker identification [47]. Although coarticulation strategies could be individualistic, they cannot be arbitrarily different. It would be interesting to factor the coarticulation, in both acoustic and articulatory domains, into one component, which is present irrespective of the speaker and another component, which varies from one speaker to another. This could be done at different speaking rates to investigate how each speaker alters coarticulation strategies at different speaking rates while achieving the acoustic and articulatory goals.

E. Acoustic-articulatory analysis of Toda front rounded vowel

Apart from the ten-vowel system, there are a number of Toda language specific vowels [48]. Among those we have recorded front rounded vowel /y/. There have been several works by Emeneau on the Toda [49], [50], where acoustic analysis of these Toda specific vowels have been carried out. The database in this paper would, for the first time, give an opportunity to carry out articulatory analysis of the Toda front rounded vowel /y/ and how it varies across speakers. That would add value to the documentation of this under documented language.

F. Other potential studies

A number of other studies can use this rich corpus. For example, how vowel onset time changes with speaking rate can be studied using the database reported in this work. How consonantal effect over vowel formants changes with speaking rate can also be studied with this database. With the help of articulatory data, one can also study oral closure patterns with respect to the labial stop consonants.

V. CONCLUSION

This work presents a database consisting of recordings from six Toda speakers while speaking thirty $V_1pV_2$ nonsense words and 42 Toda words at four different speaking rates. A unique feature of this database is that simultaneous acoustic and articulatory recordings are available to study the movements of lips, jaw and tongue. A number of research studies could benefit from this database including the effect of speaking rate and style on coarticulation, labial kinematics during stop consonants as a function of speaking rates and acoustic-articulatory analysis of Toda front rounded vowels. In future, we plan to extend the database by collecting more nonsense words with different consonants and more real-world stimuli. This would make the dataset richer allowing many more research studies to be carried out.
Acknowledgment: We would like to thank the Toda speakers for their participation in this data collection. We would also like to thank Pratiksha Trust for their support.

REFERENCES