An Analysis of Articulatory-Acoustic Data based on Articulatory Strokes

Tsuneo Kato[1,2], Sungbok Lee[1] and Shrikanth Narayanan[1]

[1] Signal Analysis and Interpretation Laboratory, http://sail.usc.edu
Viterbi School of Engineering, University of Southern California, USA
[2] KDDI R&D Laboratories Inc., http://www.kddilabs.jp, Japan
tsuneo.kato@sipi.usc.edu, sungbokl@usc.edu, shri@sipi.usc.edu

Abstract

An articulatory gestural unit representation called "articulatory stroke" is introduced. It aims to capture the constriction formation and release of a speech articulator such as the tongue tip. In that regard, the articulatory stroke is an attempt at a practical realization of the abstract articulatory gestures central to Articulatory Phonology. In this study we focus on the articulatory strokes associated with the critical articulator that is essential to realize a target phone. The critical articulatory stroke is parameterized in order to investigate the predictability of the parameters from phonetic contexts and to check the statistical dependency of acoustic changes associated with the critical articulatory strokes. Canonical correlation analysis between the articulatory strokes and ΔMFCCs showed that the critical articulatory strokes are more responsible to the acoustic changes inside target phonemes than non-critical articulators. This implies that modeling acoustic changes due to critical articulations could provide an edge in improving ASR performance.

Index Terms: articulatory stroke, critical articulation, canonical correlation analysis, automatic speech recognition

1. Introduction

Current automatic speech recognition (ASR) systems are based on surface matching between acoustic features and phonetic units comprising words in the lexicon. Though they perform quite well for read or constrained speech, they suffer from swelling acoustic variability such as when speaking style becomes spontaneous or casual, and as the number of speakers increases. It requires massive, if not endless, training data to deal with such as intra- and inter-speaker variability problems.

Modeling of articulatory movements in the articulatory domain has been considered as a promising direction to resolve the aforementioned acoustic variability problems [1, 2]. Multi-tiered articulatory description of coarticulation has been considered more suitable than the conventional segmental phonetic representation, especially for spontaneous speech. For instance, Articulatory Phonology (AP) [3] argues that coarticulation can be described more naturally by combining articulatory units called "articulatory gestures," and that syllables or words can be represented by multiple articulatory gestures with their relative timings. Furthermore, directly capturing the continuous and smooth movements of articulators toward their target positions should provide reasonable constraints on acoustic changes, which in turn can potentially offer additional means for improving ASR performance.

Acoustic signals corresponding to the articulatory movements were initially used in the landmark based approach [4]. Later, a multi-tiered articulatory configuration has been explicitly modeled within a DBN framework [5, 6] and has shown the capability of effectively representing pronunciation variations [6]. The asynchronously-evolving acoustic features and articulatory features have also been modeled by HMM/BN framework [8]. Automatic detection of gestural events has been examined with promising results [16]. Switching state space models and linear dynamic models (LDM) have been used to take the property of smooth movement toward a target position into account, both explicitly and implicitly, for decoding speech [9, 10].

Based on the aforementioned rationale, we investigate an articulatory segmental unit called "articulatory stroke" and its properties. An articulatory stroke is defined as a turnaround motion capturing the approach to, formation of, and release from a constriction. The stroke is segmented between two successive minimum curvature points. We consider the stroke as a practical realization of the articulatory gestures, which are the atomic units of speech production in Articulatory Phonology (AP) [3] framework. Most previous studies have modeled articulatory movements in combinations of discrete static states of each articulator, or in continuous states with acoustic phonetic units. However, estimating complete articulatory positions is challenging. In this work as a first step hence we focus on the motions of critical articulators, which are fundamental for producing particular sounds and closely related to the articulatory gestures. According to the sounds, the strokes are asynchronous with the perceived acoustic events. For example, the stroke forming a stop consonant occurs earlier than the resulting acoustic signature. We hypothesize that acoustic changes corresponding to the motions of the critical articulators can be more accurately captured by setting strokes as the units for ASR.

In this paper we present the results of quantitative articulatory data analyses based on the concept of articulatory strokes using the MOCHA-TIMIT database [11]. In Section 2, the articulatory stroke is defined in accordance with results of velocity-curvature analysis of the articulatory movement, and critical strokes are automatically extracted from the estimated strokes in reference to acoustic phonetic segmentation and a gestural dictionary. In Section 3, predictability of stroke parameters by parametric representation of phonetic context is examined. In Section 4, statistical dependency of acoustic feature change with each articulator motion is compared by canonical correlation analysis showing the critical articulator has higher correlation than other non-critical articulators.

2. Articulatory Stroke

2.1. Velocity and curvature analysis of articulatory movement

To find a reliable segmental unit of articulatory trajectories corroborated by its geometric and kinematic property, we investigated the velocity, acceleration and curvature properties of each sensor of the MOCHA data where the sensors are placed on the tongue tip (TT), tongue body (TB), tongue

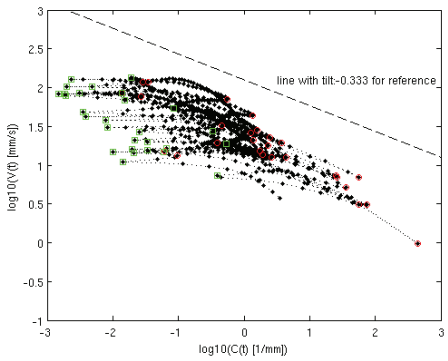


Figure 1: *A trace of log-curvature and log-velocity of the tongue tip sensor for an utterance.*

dorsum (TD), lower and upper lip (LL&UL), lower and upper incisor (LI&UI) and velum (VL). The time series data of curvature show sharp peaks, and most of their timings exactly match those of local minima of velocity and of local maxima of acceleration. This shows that the articulators make quick turns and change direction of movement in short time with their velocity slowing down.

It has been known that the velocity and curvature of articulators are roughly related by a power function called "1/3 power law", which has been observed in various motions within limb and oculomotor systems [12]. The 1/3 power law expresses the relation between velocity $V(t)$ and curvature $C(t)$ in the following equation with velocity gain factor K.

$$V(t) = KC(t)^{-1/3} \qquad (1)$$

Figure 1 shows an example of traced curvature and velocity of the tongue tip sensor for an utterance in a double logarithmic chart. The rectangular and circular markers indicate local maxima and local minima of curvature respectively. The exponent term and velocity gain factor K of the 1/3 power law are shown as an instantaneous tilt and a y-intercept of the traced curve. In making turns, curvature and velocity are well ruled by the 1/3 power law with a nearly constant K value. Though the articulator exhibits a spatially complex trajectory, it is possible to assume the movement is composed of alternate turnarounds ruled by the 1/3 power law and relatively rectilinear movements with nearly constant velocities. All sensors showed the same trend, though the curvature and velocity ranged differently. Interestingly, we found that these trends were preserved with speech style variations [15].

2.2. Extraction of critical strokes

Based on the curvature and velocity analysis, an articulatory trajectory was segmented into a sequence of units that we call "articulatory strokes". To capture a sequential motion of an articulator representing the approach to, formation of, and release from a constriction, a stroke was defined as a motion segmented by two successive points of local minimum curvature. An articulatory stroke is schematically depicted in Figure 2. It is as a turnaround motion where the curvature and velocity are ruled by the 1/3 power law with nearly constant K value. This motion captures the whole process of constriction formation: "approach" (onset), "form" and "release" (offset). However, the strokes segmented just by points of local minimum curvature contain both critical ones, which are essential for producing particular sounds, and non-critical ones that do not contribute to producing a given sound. As it was difficult to distinguish the two only by geometric property

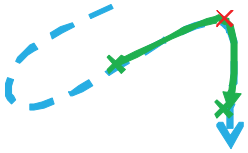


Figure 2: *Definition of an articulatory stroke: capturing the approach to, formation of, and release from a constriction. It is segmented by two successive minimum curvature points.*

without phonetic information, we referred to acoustically-obtained phonetic segmentation and the gestural dictionary which defines critical articulators for each phone [13].

The binary separation was processed in the following steps. First, strokes whose timing of local maximum curvature is within a phonetic segment for which the gestural dictionary defines a critical articulation were selected as possible critical strokes. The phonetic segments that seek a critical stroke were extended back by 20 ms because the strokes sometimes occur before the phonetic segments start. Second, strokes whose maximum curvature was under a threshold of 10 mm^{-1} at 500 Hz sampling rate, strokes whose duration was below 40 ms, and strokes whose turning point was more than two sigmas apart from the mean turning position (i.e. centroid of the turning points) of the phone were eliminated from the possible critical strokes estimation. Third, a stroke whose turning point was nearest to the mean turning position was selected as the critical one if more than two strokes exist in a phonetic segment.

2.3. Detection accuracy of critical strokes

Critical strokes are sometimes not detected in phonetic segments where the gestural dictionary assigns a gesture. Table 1 shows the rates of detected critical strokes to the numbers of phonetic segments. The detection rates were around 80% on average for all articulators. Meanwhile, we consider that the stroke concept is able to capture assimilation of critical articulation. When two successive phones have a same constriction location of a same articulator, the skipping rates of critical strokes for either of the phones were 68% for tongue tip, 84% for tongue body and 90% for lower lip. These skipping rates were much higher than the chance level based on the average detection rates for the articulators.

Table 1. *Detection rates (%) of critical strokes to the numbers of phonetic segments for consonants.*

Tongue tip	**Det. rate**	**Tongue body**	**Det. rate**	**Tongue dorsum**	**Det. rate**
ch	94.9	ch	96.0	g	95.4
d	75.0	g	83.0	k	92.2
dh	74.3	jh	88.3	ng	96.4
jh	91.3	k	85.8	average	93.6
l	76.8	l	71.9	**Low lip**	
n	80.0	ng	89.1	b	86.7
r	59.4	s	89.8	f	84.1
s	91.3	sh	93.9	m	87.4
sh	95.2	w	68.2	p	91.2
t	74.3	y	80.6	r	62.9
th	86.1	z	82.0	v	69.2
z	90.5	zh	84.0	w	84.1
zh	100.0	average	82.7	y	67.0
average	80.0			average	78.5

3. Predictability of stroke parameters from contextual parameters

3.1. Canonical correlation analysis between stroke parameters and contextual parameters

To characterize the strokes by a small number of parameters, a stroke was approximated by two piecewise lines. One line approximating the approaching part was obtained by least squares estimation (LSE) for the data points from the starting point of the stroke to the turning point, while the other line approximating the releasing part was obtained by LSE for the data points from the turning point to the end point. The turning point represented by a displacement vector from the mean turning position of the phone, approaching and releasing direction angles, distances, durations and averaged speeds were extracted as stroke parameters. Some of the stroke parameters are considered important to distinguish a particular sound in a certain context from others. The important parameters should have clear dependency with phonetic contexts. Otherwise, the parameters are deemed to carry no information on the estimation of specific phonetic contexts.

Though the strokes of a same phone can have turning points with a large variability, the mean turning position of each phone is located exactly at the place of articulation of the conventional phonetics in the midsagittal plane. We adopted a simple hypothesis that the stroke basically traces two piecewise lines, one connecting mean turning positions of the left context phone (L) and of the center phone (C), and the other connecting mean turning positions of the center phone (C) and of the right context phone (R) [14]. Therefore, the left and right phonetic contexts were represented by relative position vectors of their mean turning positions to that of the center phone $\overrightarrow{LC}$ and $\overrightarrow{CR}$ in the articulatory plane. The angles and distances of the relative position vectors were treated as hypothetical direction angles and distances of contextual parameters.

We conducted correlation analysis for all combinations of the stroke parameters of the critical strokes and the contextual parameters. Among them, two combinations of the hypothetical approaching angle as a contextual parameter and the actual approaching angle as a stroke parameter, and the hypothetical releasing angle and the actual releasing angle had higher correlation coefficients than others. The correlation coefficients of the approach and release were 0.40 and 0.40 for tongue tip, 0.23 and 0.29 for tongue body, 0.06 and 0.44 for tongue dorsum, 0.23 and 0.25 for upper lip, 0.27 and 0.25 for lower lip. In contrast, the hypothetical approaching and releasing angles and the actual turning point represented by a displacement from the mean turning position of the phone were not so correlated, with coefficients less than 0.2.

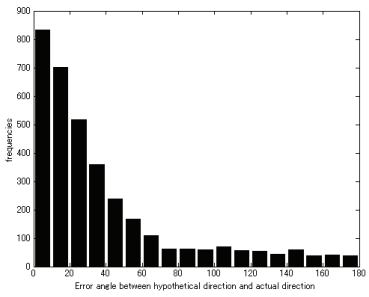


Figure 3: *Distribution of angle error between hypothetical and actual direction angles of approach part of tongue tip strokes.*

Table 2: *Canonical correlation coefficients between tongue tip (a critical articulator) motions* Δp *and* corresponding ΔMFCCs *for the approach and release from/to the turning points of a consonant /t/ with various* Δt.

/t/	Δt=10ms	Δt=20ms	Δt=30ms	Δt=40ms
approach	0.54	0.75	0.77	0.76
release	0.47	0.57	0.62	0.64

3.2. Predictability of approaching and releasing directions of strokes

The predictability of approaching and releasing direction angles of strokes was evaluated by the angle error between the hypothetical and actual approaching (releasing) angles. Figure 3 shows the distribution of angle error between the hypothetical and actual approaching angles of the critical strokes of tongue tip. The angle error was below 30 degree for 60% of the strokes. The accuracy should be improved. In this simple prediction, we did not consider whether the mean turning positions of the left and right context phones were of the critical or dependent or redundant articulators [17]. The predictability is expected to be improved by taking the types of the articulator into account.

4. Canonical correlation between articulatory motion and acoustic feature change

4.1. Canonical correlation analysis

Hypothesizing that the articulatory gestures are the atomic units of speech perception, acoustic features corresponding to articulatory strokes should be observed in the speech signal. We hypothesized that the critical articulator is most dynamic among all articulators, and that acoustic feature change is most correlated with the motion of the critical articulator. As we do not know how each articulator's motion affects the acoustic features, we performed canonical correlation analysis between motions of articulatory strokes and corresponding MFCCs changes with optimized time window for delta features, and compared them between articulators. The base acoustic features were 12-order MFCCs with 25 ms window and 10 ms frame shift.

4.2. Time window for delta

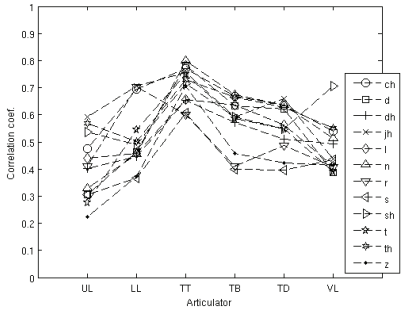
Let t, $p(t)$ and Δt be the instant of turning point of a critical articulator, its position (x, y) in the articulatory plane and the time window to get deltas, the approaching and releasing motions of the critical articulator are defined as follows:

$$\Delta p_{app}(t) = \frac{p(t) - p(t-\Delta t)}{\Delta t} \quad (2)$$

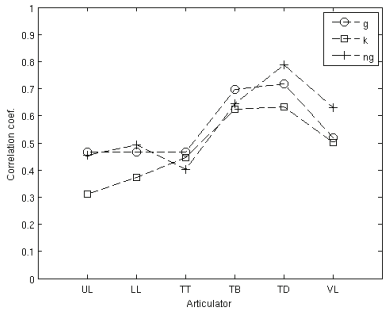
$$\Delta p_{rel}(t) = \frac{p(t+\Delta t) - p(t)}{\Delta t} \quad (3)$$

The acoustic feature changes corresponding to the approaching and releasing motions of the critical articulator are $\Delta MFCCs$ defined in the same way as equations (2) and (3) with using $MFCC(t)$ instead of $p(t)$.

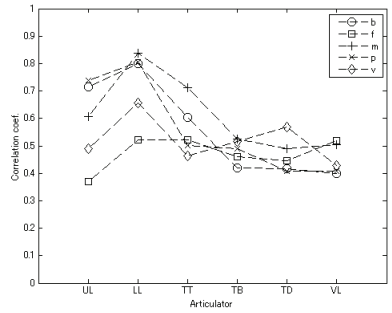
The correlation between articulatory motions Δp and $\Delta MFCCs$ are expected to be higher with Δt of the stroke duration, which is longer than a 10 ms frame shift, because it is less affected by artifacts even if the temporal resolution comes down. To find an optimal time window size for delta, the correlation coefficients between Δp_{app} and $\Delta MFCC_{app}$ and Δp_{rel} and $\Delta MFCC_{rel}$ were compared experimentally across various values of Δt.



A: Consonants for which tongue tip is the critical art.



B: Consonants for which tongue dorsum is the critical art.



C: Consonants for which lips are the critical articulators.

Figure 4: *Correlation coefficients between articulatory motions of each articulator and the corresponding ΔMFCCs.*

Table 2 shows the correlation coefficients of the consonant /t/ by a speaker in MOCHA. The critical articulator was specified using the canonical rule mentioned earlier. The correlation coefficients reached their maxima at Δt of 30 ms and the maximum values were higher than those of 10 ms case.

However, Δt which gives the maximum values is shorter than the mean stroke duration. This implies that the critical articulator is not dominant for the acoustic change for the entire duration from the beginning to the end of a stroke.

4.3. Comparison between articulators

Figure 4 shows canonical correlation coefficients between Δp of each articulator and *ΔMFCCs* for each consonant. The three panels show results for consonants for which tongue tip (panel A), tongue dorsum (panel B) and lips (for panel C) are the critical articulators respectively. Δt was set at 30 ms. Basically, the correlation coefficient of the critical articulator is higher than those of other non-critical articulators as expected. Only in panel C, the correlation coefficient of lower lip is higher than that of upper lip, because lower lip is innately more dynamic than the upper lip. Therefore, the critical strokes of approaching and releasing are the dominant factor of the corresponding ΔMFCC.

5. Conclusions

Motivated by joint modeling of articulatory and acoustic streams, the articulatory trajectories of MOCHA database were segmented into units of "articulatory strokes" in an articulator dependent fashion. We consider the stroke as an implementation of constriction formation by a critical articulator. With reference to the available acoustic phonetic segmentation and a gestural dictionary, a stroke was detected for around 80% of the phonetic segments of the articulator for which the gestural dictionary assigns articulatory gestures.

The actual approaching and releasing directions of the strokes were correlated with the hypothetical approaching and releasing directions of the lines connecting the mean turning positions of the successive phones, left context, center phone and right context. The angle error of the prediction was below 30 degree for 60% of the strokes.

As to statistical dependency of motions of the strokes and acoustic feature change, canonical correlation analysis between the motions and ΔMFCC showed that the approaching and releasing motions of the critical articulator were more correlated with corresponding ΔMFCC than other non-critical articulators. This result means that the motions of the critical articulators are the most dynamic and the dominant factor of acoustic feature change. We are currently working on automatic detection and meaningful classification of the critical strokes with only their geometric property, and on improving the stroke parameter predictability using the classification of critical, dependent and redundant articulators as the second step.

6. References

[1] S. King et al. "Speech production knowledge in automatic speech recognition," J. Acoust. Soc. Am., Vol.121 (2) pp.723-742 (2007)
[2] E. McDermott et al. "Production-oriented models for speech recognition," IEICE Trans. Inf.&Syst., Vol.E89-D, No.3, pp.1006-1014 (2006)
[3] C. P. Browman and L. Goldstein, "Articulatory phonology: an overview," Phonetica, Vol. 49(3-4), pp. 155-180 (1992)
[4] K. N. Stevens et al. "Implementation of a model for lexical access based on features," Proc. of ICSLP1992 (1992)
[5] J. Sun, X. Jing, L. Deng, "Data-driven model construction for continuous speech recognition using overlapping articulatory features," Proc. of ICSLP 2000 (2000)
[6] K. Livescu and J. Glass, "Feature-based pronunciation modeling with trainable asynchrony probabilities," Proc. of ICSLP 04, (2004)
[7] K. Livescu and J. Bilmes, "Hidden feature models for speech recognition using dynamic bayesian networks," Proc. of Eurospeech 03, pp.2529-2532 (2003)
[8] K. Markov and S. Nakamura, "Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework," Speech Comm., Vol.48, pp.161-175, (2006)
[9] L. Deng, "Switching dynamic system models for speech articulation and acoustics," Proc. of IMA workshop, (2000)
[10] J. Frankel, "Linear dynamic models for automatic speech recognition," Ph.D. thesis, University of Edinburgh (2003)
[11] A. Wrench, "A new source for production modeling in speech technology," Proc. of Workshop on Innovations in Speech Processing (2001)
[12] S. M. Task and J. R. Westbury, "Speed-curvature relations for speech-related articulatory movement," J. of Phonetics, Vol.32, pp.65-80 (2004)
[13] "TADA manual v0.9", http://www.haskins.yale.edu/
[14] C. S. Blackburn and S. Young, "A self-learning predictive model of articulator movements during speech production," J. Acoust. Soc. Am., Vol.107, No.3 , pp.1659-1670 (2000).
[15] S. Lee, T. Kato,and S. Narayanan, "Relation between geometry and kinematics of articulatory trajectory associated with emotional speech production," Proc. of Interspeech 08 (2008)
[16] A. Gutkin and S. King, "Detection of symbolic gestural events in articulatory data for use in structural representations of continuous speech," Proc. of ICASSP 05 (2005)
[17] V. D. Singampalli and P. Jackson, "Statistical identification of critical, dependent and redundant articulators," Proc. of Interspeech 07 (2007)