Prosodic Timing Analysis for Articulatory Re-synthesis Using a Bank of Resonators with an Adaptive Oscillator

Michael C. Brady

Department of Cognitive and Neural Systems, Boston University, USA

mcbrady@bu.edu

Abstract

A method for the analysis of prosodic-level temporal structure is introduced. The method is based on measured phase angles of an oscillator as that oscillator is made to synchronize with reference points in a signal. Reference points are the predicted peaks of acoustic change as determined by the output of a bank of tuned resonators. A framework for articulatory re-synthesis is then described. Jaw movements of a robotic vocal tract are made to replicate the mean phase portrait of an utterance with reference to a production oscillator. These jaw movements are modeled to inform the dynamics of within-syllable phonemic articulations.

Index Terms: suprasegmental timing analysis, dynamic time warping, articulatory synthesis, speech robotics.

1. Introduction

Dynamic time warping [1, 2] is a technique used in artificial speech recognition and processing. The technique stretches and/or compresses a piece of recorded speech to some standard duration for statistical comparison with other time-normalized samples. Important information for mapping a speech segment to a standard duration comes with reference to the temporal structure of surrounding speech. In other words, an estimate of speech tempo is useful.

Numerous studies on speech timing and attention indicate that production-perception centers or (P)-centers seem to hold a special importance during mental integration through time of the speech signal [3-5]. (P)-centers generally correspond to vowel onset (VO) regions of the speech stream and VO locations are convenient to estimate. VOs are taken as points of maximal positive change in the low-pass amplitude envelope of a recorded signal as depicted in Figure 1. As will be discussed, (P)-centers also seem to relate to anticipatory movements of the jaw. A method based on circular statistics that uses predicted VOs as reference points for inferring a speech tempo is described and evaluated.

2. Method

In circular statistics there are two variables to be considered. One is the recurring event of interest and the other is the period of a reference sinusoid. For instance, we can measure how reliably or with what variance a train arrives on the hour with reference to the hour hand on a clock. For a review of circular statistics, see [6, 7]. In circular statistics as applied to speech tempo, we must work a bit backwards. The timing of the recurring event (the VO) is known and the period of the reference sinusoid needs to be determined. Once an optimal period is estimated, we may analyze how phase measurements of VOs cluster. Recorded utterances that exhibit highly regular or metrical timing will result in strong VO clustering and will perhaps exhibit an occasional *discordant VO*. The discordant VO relates to the outlier in normal statistics.

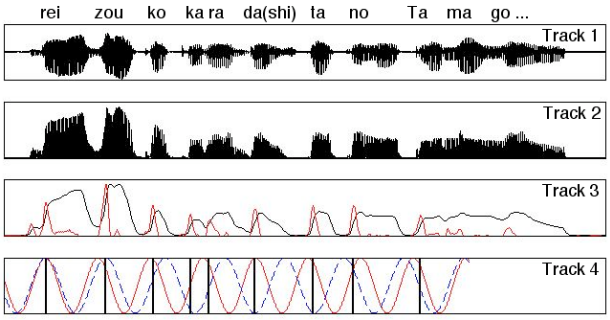

Figure 1: *estimating VO phase angles from a signal. The raw signal (Track 1) is low-pass filtered and full rectified to return the amplitude envelope (Track 2). The positive change in slope (derivative) of a smooth-filtered version of the envelope is used to mark areas of maximal change in the signal (Track 3). These areas generally correspond to the onsets of vowels. A peak-finding algorithm with threshold is then used to approximate the locations of onsets at discrete points in time (Track 4). Where these discrete locations intersect with two example reference or 'base' sinusoids define two sets of phase measurements.*

2.1. Phase clustering and the base sinusoid

Figure 1 depicts how VOs are extracted for a recording of the Japanese utterance "reizouko kara dashita no tamago," ('the egg taken from the refrigerator'). Phase measurements for these VOs are realized from two sinusoids with different periods. The period of one sinusoid (solid) is 192ms while the period of the other sinusoid (dashed) is 218ms. VO phase measurements for each sinusoid are plotted circularly in Figure 2. The VOs are labeled in terms of the consonant-vowel pair associated with each VO. Notice that the phase circle corresponding to the sinusoid with a period of 192ms (left) exhibits relatively strong VO clustering while the phase circle corresponding to the 218ms sinusoid (right) exhibits virtually no clustering. Degree of clustering is quantified with a measure based on R, the length of the sum of phase angle

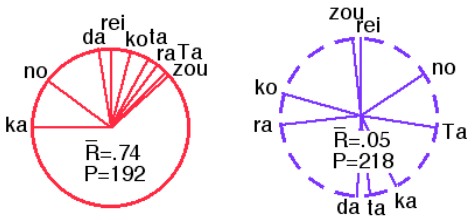

Figure 2: *phase clustering. Phase measurements taken with respect to the solid sinusoid (left) and dashed sinusoid (right) from Track 4 of Figure 1 are plotted. Clustering for each is measured in terms of* $\bar{R}$.

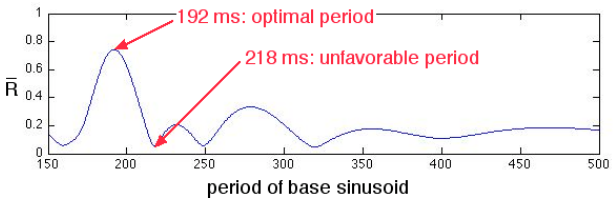

Figure 3: $\bar{R}$ *values of the VO sequence when sweeping through a spectrum of base sinusoids with periods between 150ms and 500ms. The optimal and unfavorable periods seen in Track 4 of Figure 1 and used to generate Figure 2 are indicated.*

vectors as expressed in Eq. 1, where θ_i is the phase of a VO. The mean resultant length $\bar{R}$ is scaled to a range between 0.0 and 1.0 based on *n,* the number of measurements used. The greater the $\bar{R}$ value, more the phase angle vectors point in the same direction and thus the better the phase clustering.

$$R^2 = \left(\sum_i \sin(2\pi\theta_i)\right)^2 + \left(\sum_i \cos(2\pi\theta_i)\right)^2 \qquad \bar{R} = \frac{R}{n} \qquad (1)$$

Figure 3 plots $\bar{R}$ for the VO sequence of Figs 1 and 2 across a range of periods that might have been selected for the base sinusoid. Sweeping a VO sequence with a range of sinusoids relates to performing a Fourier transform. From this sweep, an optimal base sinusoid period is selected for circular analysis.

In inspecting the clustering of the 192ms phase circle of Figure 2 (left), there are two VOs that may be considered as discordant. That is, if the phase angles for the VOs labeled 'no' and 'ka' had not been included in the $\bar{R}$ calculation, $\bar{R}$ would have been larger. We will return to this.

2.2. Resonator bank

Speech is riddled with timing variability. For expressive and systematic reasons as well as for arbitrary ones, fluent speech deviates from being timed with integer ratios. However, it is not necessary for speech to exhibit periodicity in order to use circular analysis. We need only to map recorded speech samples to sinusoids in a consistent way to be able to compare across resulting phase portraits. It is also useful to map a sinusoid to an utterance as the utterance unfolds through time.

One way of synchronizing a sinusoid with the VOs of an utterance as the utterance arrives through time is to first employ a bank of tuned resonators to filter the VO sequence so as to emphasize any approximate periodicity the sequence may have. The resonator bank is introduced elsewhere [8, 9] and only a brief overview can be provided here. A resonator is modeled as a pendulum. A pendulum has a natural period it will swing with based on the length of its cord. Pushing a pendulum during its forward swing will increase the amplitude of the pendulum while pushing during its backswing will decrease its amplitude. When pushing a pendulum in the middle of its forward swing, swing amplitude is maximally increased and this is called the moment of the pendulum's 'maximal expectation.' Because a pendulum will synchronize its maximal expectations with the inputs of a periodic signal (if the natural period of the pendulum is the same as the period of the input signal), polling the pendulum's amplitude at its moment of maximal expectation provides information about how periodic an input signal is in respect to the pendulum. If a bank of pendulums having a range of natural periods is made to respond to the same input signal, the summed output of the bank may be used to assign temporal salience scores to the input events. Fig. 4 illustrates how the VO sequence we have been using is assigned salience scores from the resonator bank.

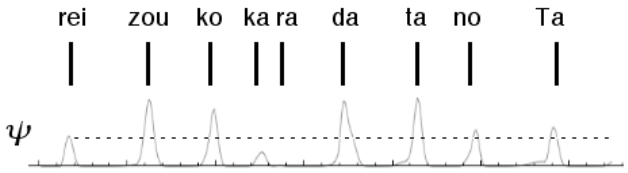

Figure 4: *temporal salience scores, ψ, assigned to the VOs of Figure 1 by the resonator bank as the VOs arrive through time. Response by the resonator bank from a resting state to the initial VO ('rei') defines a neutral score to be used as a threshold. This threshold is marked by the dashed line. Notice that the VOs 'ka' and subsequent 'ra' are below threshold while the other VOs are all at or above threshold.*

2.3. Base sinusoid as adaptive oscillator

An adaptive oscillator is made to synchronize with the VOs of an utterance through time based on salience scores passed through a threshold function, σ. If left alone, the oscillator will simply produce a sinusoid. However, at the time of a VO there may be a slight adjustment to both the oscillator's period, ρ:

$$\dot{\rho} = \rho \cdot \eta \cdot \sigma(\psi_{VO}) \cdot \sin(2\pi\theta) \qquad (2)$$

and its phase, θ:

$$\dot{\theta} = \rho \cdot \eta \cdot \sigma(\psi_{VO}) \cdot -\sin(2\pi\theta) \qquad (3)$$

This adjustment roughly follows e.g. [10, 11] and essentially corrects the oscillator for gradual speech tempo shifts. Here, η is the oscillator's coupling strength. Note that the coupling strength need not be the same value for Equations 2 and 3.

3. Analysis

For the sake of comparison, a repetition of the phrase we have been considering was analyzed. Both recordings came from a corpus provided by a JST/CREST ESP project headed by Nick Campbell. Kuroyanagi Tetsuko, a well-known Japanese talk show host and author read aloud the popular book "Usotsugu Tamago." The audio book was segmented and then notated for statistical analysis. The phrase in question occurred twice in

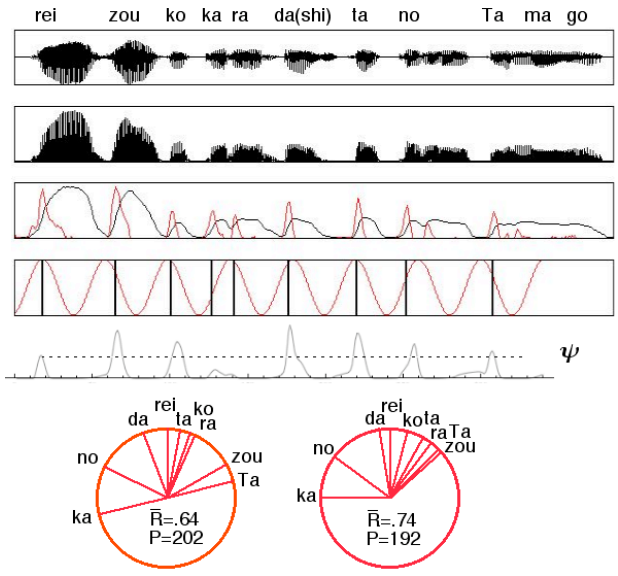

Figure 5: *analysis of a repetition of the phrase from Figure 1. The phase portrait of the optimal* $\bar{R}$ *sinusoid* ρ*=202ms for the sample (left) is compared with the optimal phase portrait of the original sample (right).*

the corpus within different sentences from different chapters. Japanese allows the method of this paper to best be illustrated using only two short samples because Japanese exhibits a certain timing regularity due to the nature of its duration-based phonological distinctions. Figure 5 shows how VOs and related information were extracted for this new recording. The resulting phase portrait with an optimal base sinusoid period of 202ms is plotted next to the phase portrait of the first sample.

3.1. Results

The adaptive oscillator was initialized with an optimal period and was made to synchronize with the peaks of the ψ signal that were above threshold following Equations 2 and 3 for each of the two utterances. The resulting sinusoidal signal was used as the base sinusoid from earlier. Results are presented in Figure 6 where Sample 1 corresponds to the first utterance and sample 2 corresponds to the repetition just introduced. It can be seen that phase clustering is greatly improved. This is due to the elimination of discordant VOs. Specifically, 'ka' was excluded (along with 'ra'), and 'no' was filtered toward the center of the cluster in terms of ψ. As is demonstrated by these examples (and other examples from various languages not described here), tracking by the adaptive oscillator through time to the ψ signal is shown to be a reliable way of estimating the tempo of an utterance and of assigning a metric from which to perform comparative circular analyses. It should be noted that there was virtually no shift in speech tempo for these two sample utterances but that in running speech there often tends to be a tempo shift e.g. at the end of the phrase.

Figure 6 also depicts a phase portrait that is the mean portrait for the two utterances. This portrait is used during re-synthesis, to be described next.

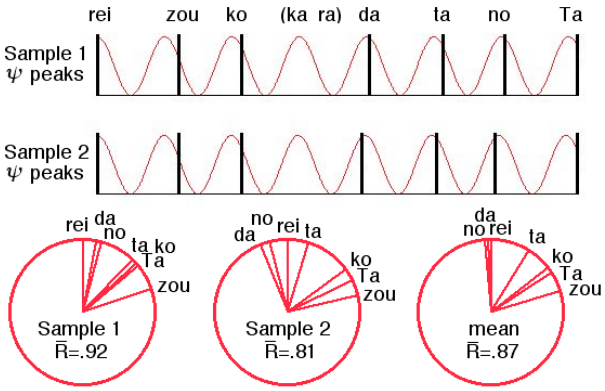

Figure 6: *improved phase clustering. By excluding the 'ka' and 'ra' VOs and by migrating the 'no' and other VOs toward the center, phase portraits of the two utterances become more periodic, as measured by* $\bar{R}$.

4. Articulatory re-synthesis

A production oscillator is used as reference for generating a stream of articulatory motor movements. The oscillator allows articulator motor dynamics to anticipate the timing of future VOs based on respective phase angles. This arguably results in a more natural sounding prosodic contour than is typically realized when sequencing phonetic segments together like "beads on a string," with no reference to global timing.

Articulatory re-synthesis (as opposed to concatenative or unit selection re-synthesis) is the focus of investigation here because the analysis method ultimately relates to a theory of speech motor control [8-9]. However, the technique may lend itself well to other re-synthesis approaches.

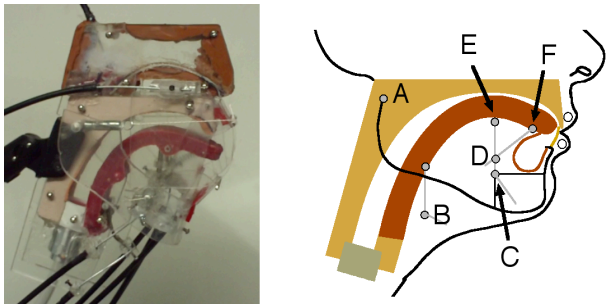

Figure 7: *mechanical vocal tract (left) and schematic for how tongue shape is manipulated (right). The jaw rotates about Point A and the tongue root moves back and forth by rotating at Point B. The body of the tongue moves back and forth by rotating at Point C and is raised and lowered by manipulating the distance between Points C and E. Point C is attached to the jaw. The tongue tip moves by manipulating the distance between Points D and F.*

4.1. Mechanical vocal tract

A mechanical vocal tract is strategic to use during re-synthesis for a variety of reasons. For one, a mechanical tract "computes the physics for free." Most digital articulatory speech synthesizers are based on a source-filter model of vocal production rather than on calculating how air particles reverberate in the vocal cavity. A truly aerodynamic model realizes the non-linear interaction between vocal cord vibration and vocal cavity shape. This difference is important, for instance, in synthesizing natural-sounding female speech. See e.g. Titze [12] for further discussion on this topic.

The mechanical vocal tract has a silicone tongue that serves as the bottom of a plexiglass and resin enclosed vocal cavity. The tongue is actuated by five degrees of freedom as depicted in Figure 7. Other degrees of freedom involve the lips, vocal cords, and velum open-close for nasals. The voicing source has been implemented using both an electric loudspeaker and using compressed air in conjunction with artificial vocal cords. Video demonstrations of the system for Japanese as well as for other languages may be found at: http://www.fluidbase.com/vocaltract

4.2. Jaw movement and prosody

Analysis of articulatory data indicates that movement of the jaw during fluent speech may generally correspond to VOs and (P)-centers [4]. It has been shown that amount of jaw articulation (excursion) relates to prosody in terms of lexical stress, phrasal prominence, and syllable position within the phrase [e.g. 13]. It has also been shown that the jaw opens more during increased phrasal prominence for low, mid, and high vowels [14]. It is thus theorized that jaw control sets the framework for prominence characteristics of the syllable [15]. In contrast to jaw movement, tongue movement seems more to be controlled appropriately for the production of the desired vowel given local articulator conditions. Peak F0 and acoustic syllable duration within the utterance also seem to correspond to local articulator conditions. It is thus reasonable to hypothesize that jaw movements correspond to the timings associated with prosody while tongue and ongoing vocal cues may be viewed more as residual to prosodic control.

Recent work [16] indicates patterns of alternating amounts of jaw displacement in English. In producing a series of monosyllabic words with the same vowel, jaw displacement of a speaker was shown to vary systematically between 'strong' and 'weak.' Another study [17] found the timing

locations of jaw excursions as well as phrase boundaries from articulatory data, including lip and tongue articulation of the syllable. It was determined that timing as well as location and size of phrase boundaries do not change the inherent rhythm of an alternating strong-weak jaw displacement. In light of these and other studies, an articulatory model for alternating strong-weak jaw displacement may be called for.

4.3. Results

Both jaw movements and voicing onsets were made to replicate the mean VO phase portrait of the phrase we have been considering. While running the production oscillator at 4Hz (a period of 250ms), the jaw was made to open so as to reach its peak displacement each time the production oscillator reached the phase of its next target VO. Following each VO, the jaw was then made to decay towards a neutral position. The tongue articulators were algorithmically made to compensate for the change in jaw position as they sought to produce the vowel changes associated with the VOs. In an attempt to model possible alternating jaw displacement for Japanese, two more utterances were re-synthesized where the jaw was made to alternate in degree of openness for target VOs. One had {'rei,' 'ko,' 'da,' 'no'} as emphasized while the other had {'zou,' ('ka,' 'ra'), 'ta,' 'Ta'} as emphasized.

Analysis of the re-synthesis method and its results can only be qualitative given the short nature of this paper. By definition, the mean phase portrait of Figure 6 is precisely replicated in the re-synthesized utterances and there is no need to provide a Figure for this. For a better idea of results, the reader is encouraged to inspect videos of the re-synthesized utterances by the mechanical vocal tract (as well as the audio clips used in the speech analysis section of this paper). This media is made available on the mechanical vocal tract's web site. Though the tract cannot yet produce good consonants, subjective assessment of the re-synthesized phrases finds them to capture a naturalistic prosody for the utterance. The videos of alternating strong-weak jaw movements for the phrases provides perceptual contrast in appreciating the re-synthesis method, but further work is needed before any claims can be made about alternating jaw excursion in Japanese.

5. Discussion

A variety of motor theories of speech production stand to benefit from the suprasegmental timing analysis and jaw-oscillator re-synthesis approach presented here. For instance, the Theory of Articulatory Phonology [18] posits that the speech gestures of an utterance are triggered by internal oscillators, where the motor 'plan' is a graph that couples gestural oscillators with one another. Such a description is integral to the notion that the speech signal can be analyzed with reference to an utterance-level timing metric. The DIVA model of speech motor learning and control [19, 20] motivates this work. DIVA models speech perception and production from a neuroscience framework. Appreciating the neural substrates of speech timing and sequencing is a very active research area for DIVA. Modeling the mechanisms involved with the fluent production of speech may, for example, help in treating people with stuttering disorders.

The idea of an alternating strong-weak jaw displacement in Japanese is an open question. Though Japanese does not exhibit alternating strong-weak syllables as does e.g. English or Russian, the notion of a bimoraic timing and the foot in Japanese lends itself to the analysis for alternating jaw movement with respect to the relatively fast optimal analysis tempos of Japanese. In English, the periodicity of accented syllables might well be mapped to the quarter note or to what is known as the tactus pulse of Western music, typically in the range of 2Hz. This periodicity is about double that of the periods of the analysis and re-synthesis sinusoids of this paper.

There are a number of issues yet to be discussed. For example, how reliably do jaw movements align with the VOs of the acoustic signal in Japanese and other languages? Is the concept of an underlying speech tempo applicable to all languages, and, if so, how do the speech tempos of various languages compare with one another? Other issues have to do with tuning the parameters of the system. What are the optimal values for coupling strength in Equations 2 and 3, how much should the jaw move depending on vowel category, and how might VOs be extracted from the auditory signal when the low-pass envelope remains relatively constant? How might the success of this approach be best evaluated? These and other questions are addressed in continuation of this work.

6. Acknowledgements

Support comes from NIH grants R01 DC002852 and R01 DC007683. Thanks go to Frank Guenther, Donna Erickson, Robert Port, and three anonymous reviewers.

7. References

[1] Rabiner, Rosenberg, & Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," IEEE Trans. Acoustics, Speech, and Signal Proc., Vol. 26, 1978.

[2] Sakoe, H. and Chiba, S. "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. on Acoustics, Speech and Signal Proc., Vol. 26, 1978.

[3] Allen, G., "Speech rhythm: Its relation to performance universals and articulatory timing," J. of Phonetics 3, 1975.

[4] de Jong, K. "The correlation of P-center adjustments with articulatory and acoustic events," Perception and Psychophysics, 56 (4) pp. 447-460, 1994.

[5] Port R., "Meter and speech," J. of Phonetics 31, 599-611, 2003.

[6] Fisher, N. I., "Statistical Analysis of Circular Data," Cambridge University Press, 1993.

[7] Brady, M.C., & Port, R.F., "Quantifying vowel onset periodicity in Japanese", Proc. 16th ICPhS, 2007.

[8] Brady, M.C., "Adaptive resonance situated for articulatory speech learning and synthesis," Proc. ICDL, 2006.

[9] Brady, M.C., "Speech as a problem of motor control in robotics," Proc. 31st Meeting of Cog. Sci. Society, 2009.

[10] Large & E., Jones, M., "The dynamics of attending: How we track time varying events." Psych. Review 106, 119–159, 1999.

[11] Barbosa, P., "How prosodic variability can be handled by a dynamical speech rhythm model," Proc. 16th ICPhS, 2007.

[12] Titze, I., "Nonlinear source-filter coupling in phonation: theory," J. Acoustic Society of America, 123(5), 2008.

[13] Erickson, D., "Effects of contrastive emphasis on jaw opening," Phonetica, 55, 147-169, 1998.

[14] Erickson, D., "Articulation of extreme formant patterns for emphasized vowels," Phonetica, 59, 134-149, 2002.

[15] Fujimura, O., "The C/D model and prosodic control of articulatory behavior," Phonetica 57, 128-138, 2000.

[16] Erickson, D., "An articulatory account of rhythm, prominence, and articulatory organization," Speech Prosody, 2010.

[17] Bonaventura, P. & Fujimura, O. "Articulatory movements and phrase boundaries," in: P. Beddor, J. Ohala and M. Solé [Eds.], Experimental Approaches to Phonology, Oxford University Press, Oxford, 2007.

[18] Goldstein, L., Bryd, D., Saltzman, E. "The role of vocal tract gestural action units in understanding the evolution of phonology," in: Arbib, M. [Ed.] Action to Language via the Mirror Neuron System. Cambridge University Press, 2006.

[19] Guenther, F.H., "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production," Psychological Review, 102. pp 594-621, 1995.

[20] Tourville, J.A., Reilly, K.J., and Guenther, F.H. "Neural mechanisms underlying auditory feedback control of speech," NeuroImage, 39, pp 1429-1443, 2008.