Journal of Phonetics 65 (2017) 126–144

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

# The role of time in phonetic spaces: Temporal resolution in Cantonese tone perception

Kristine M. Yu

*Department of Linguistics, University of Massachusetts Amherst, Amherst, MA 01003, United States*

ARTICLE INFO

*Article history:*
Received 14 June 2016
Received in revised form 14 June 2017
Accepted 25 June 2017

*Keywords:*
Tone perception
Tone recognition
Temporal integration
Temporal resolution
Cantonese
Interrupted speech

ABSTRACT

The role of temporal resolution in speech perception (e.g. whether tones are parameterized with fundamental frequency sampled every 10 ms, or just twice in the syllable) is sometimes overlooked, and the temporal resolution relevant for tonal perception is still an open question. The choice of temporal resolution matters because how we understand the recognition, dispersion, and learning of phonetic categories is entirely predicated on what parameters we use to define the phonetic space that they lie in. Here, we present a tonal perception experiment in Cantonese where we used interrupted speech in trisyllabic stimuli to study the effect of temporal resolution on human tonal identification. We also performed acoustic classification of the stimuli with support vector machines. Our results show that just a few samples per syllable are enough for humans and machines to classify Cantonese tones with reasonable accuracy, without much difference in performance from having the full speech signal available. The confusion patterns and machine classification results suggest that loss of detailed information about the temporal alignment and shape of fundamental frequency contours was a major cause of decreasing accuracy as resolution decreased. Moreover, machine classification experiments show that for accurate identification of rising tones in Cantonese, it is crucial to extend the temporal window for sampling to the following syllable, due to peak delay.

© 2017 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

## 1. Introduction

A central goal of phonetics is understanding what parameters are relevant for defining phonetic categories[1] (Ladefoged, 1980). If we define a phonetic parameter to be any property relevant to the human determination of speech sounds, then these parameters include contributions from visual (e.g. McGurk & MacDonald, 1976) and top-down contextual (e.g. Ganong III, 1980) properties. Narrowly speaking, though, phonetic parameters are often thought as acoustic, articulatory, or auditory, and this subset of parameters is what we focus on here.

A careful understanding of the relevant parameters for a phonetic category, e.g. the high tone ˥ in utterance-initial position preceding a mid tone ˧ in Cantonese, or the [ae] vowel in English, is crucial: *how we understand the recognition, dispersion, and learning of a set of phonetic categories is entirely predicated on what parameters we use to define the phonetic space that they lie in*. A natural way to sharpen what we mean by "relevant" parameters is to ask: what are the *primary*[2] measures or parameters that distinguish among a set of phonetic categories? For example, one familiar set of primary acoustic parameters for defining English vowels is the first and second formants (measured at steady state) (Peterson & Barney, 1952).

The high-level contribution of this paper is to draw attention to a source of phonetic parameters which is sometimes overlooked: time. The speech signal unfolds in time, so any parameter, e.g. the first formant (F1) or fundamental frequency (f0), actually expands into a *family* of parameters, where the number of parameters in the family depends on the temporal

*E-mail address:* krisyu@linguist.umass.edu

[1] A *category* defines a density distribution over some parameter space, and a *phonetic* category is defined over a phonetic parameter space (Pierrehumbert, 2003, p. 119). This definition leaves open how a category might come into the phonology, e.g. as a contextually-conditioned allophone or a phoneme. This is an active area of research that isn't central to this paper, but see Pierrehumbert (2003), Peperkamp, Calvez, Nadal, and Dupoux (2006), Dillon, Dunbar, and Idsardi (2013) for further discussion.

[2] Even setting aside non-acoustic properties, there are already an infinite number of parameters one could extract from the speech signal, and the primary cues that listeners use to identify sound categories may be large (Lisker, 1978) and flexible (Clarke & Garrett, 2004; Liu & Samuel, 2004; Sumner & Samuel, 2009; Whalen & Xu, 1992). Thus, what one means by "primary" must also be operationalized, but no matter what the criteria, we must have some way of limiting the number of parameters in play for scientific interpretability.

http://dx.doi.org/10.1016/j.wocn.2017.06.004
0095-4470/© 2017 The Author(s). Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

resolution chosen. Each parameter is sampled over time with some resolution, e.g. if F1 is parameterized to be measured at the onset and offset of a vowel, then the F1 family contributes 2 dimensions to the defined vowel space; if absolute f0 is parameterized to be measured at 8 timepoints for a tone, then it contributes 8 dimensions to the defined tonal space.[3]

To see how the study of the recognition, dispersion, and learning of a set of phonetic categories is predicated on the choice of temporal resolution (just as it is predicated on any other phonetic parameter), consider our understanding of vowels. Automatic vowel formant extraction software (e.g. FAVE-extract)–which is ever more frequently used in this age of big data—commonly extracts formants from just a single point in the vowel (Evanini, 2011; Labov, Ash, & Boberg, 2006; Reddy & Stanford, 2015; Rosenfelder, Fruehwald, Evanini, & Yuan, 2011, 2014). This is done despite substantial evidence that formant trajectories are important for vowel identification, especially for English, with its preponderance of diphthongal vowels (Nearey & Assmann, 1986; Strange, Jenkins, & Johnson, 1983). Moreover, studies of the typology of vowel dispersion (Becker-Kristal, 2010; Liljencrants & Lindblom, 1972) have used input data consisting of formant values measured at just a single point, as have simulations for learning vowel categories from phonetic data (de Boer & Kuhl, 2003; Feldman, Griffiths, Goldwater, & Morgan, 2013; Vallabha, McClelland, Pons, Werker, & Amano, 2007). There is thus a vast body of literature on vowel recognition, dispersion, and learning that rests on the potentially tenuous assumption that the relevant temporal resolution for formant values is a single sample per vowel.

Whether we define a tonal space as 8-dimensional, with 8 f0 samples over the syllable (Khouw & Ciocca, 2007), or as 2-dimensional, with f0 samples measured at the onset and offset of voicing (Barry & Blamey, 2004), has the same far-reaching consequences. For example, Kuang (2013) showed that once phonation parameters in addition to f0 parameters are included as dimensions in a tonal space, we can understand how listeners can possibly discriminate between the five level tones of Black Miao. But choosing the temporal resolution for f0 can have as much import as deciding whether or not non-f0 properties should be included as parameters in a tonal space. Alexander (2010) found that tones were not always well-dispersed within an inventory across a range of languages if the tonal space was defined using a single f0 point; however tones were well-dispersed if they were defined in a 2-D tonal space over f0 measured at the onset and offset of the syllable. Choosing how to parameterize f0 properties also matters; Gauthier, Shi, and Xu (2007) found that neural networks learned to classify Mandarin tones with higher accuracy when the tonal space was defined over the f0 velocity timecourse than the f0 timecourse.

This paper uses Cantonese tone perception as a case study to examine the contribution of fine temporal detail to the parameterization of phonetic spaces. We use perceptual and acoustic evidence to address how humans and machines respond in tone identification as the temporal resolution in the speech signal is systematically lowered.

### 1.1. Temporal resolution for tonal concepts

The degree of fineness of temporal resolution in the speech signal relevant for human cognition is still an open question. It has long been assumed in linguistics that fine temporal resolution of the speech signal is not necessary for the parameterization of tones in tone languages, and discussions of temporal resolution have largely been confined to the automatic tonal recognition literature. Chao, who introduced the iconic tone letters (Chao, 1930) used in the International Phonetic Alphabet for representing linguistic tone, wrote: “the exact shape of the time-pitch curve, so far as I have observed, has never been a necessary distinctive feature, given the starting and ending points, or the turning point, if any, on the five-point scale” (Chao, 1968, 25), and tone letters are understood to have up to 3 samples, e.g. ˩˧˥.

Additionally, Laniran (1992) argued for two targets per tone in Yoruba, and Barry and Blamey (2004) argued for 2-D acoustic Cantonese tonal spaces defined over onset and offset f0 values based on perceptual dimensions hypothesized from multidimensional scaling analyses of cross-linguistic tonal perception (Gandour & Harshman, 1978; Gandour, 1981, 1983). Morén and Zsiga (2006) and Zsiga and Nitisaroj (2007) argued for one target per tone in the representation of Thai tones in connected speech, with falling tones associated with a peak at syllable midpoint, high tones with a peak at syllable offset, rising tones with an f0 minimum at syllable midpoint, and low tones with a low pitch target at syllable offset.

In contrast, the computational literature has sometimes presumed that much finer sampling is valuable. In a study of unsupervised learning of Mandarin tones, Gauthier et al. (2007) extracted 30 samples of f0 or 28 samples of f0 velocity per syllable (a sampling rate on the order of 1 sample every 10 ms), and a number of automatic tonal recognizers parameterize the f0 curve by sampling f0 every 10 ms, e.g. Pisarn and Theeramunkong (2006), Prukkanon, Chamnongthai, and Miyanaga (2016), Zhang and Hirose (2004). But in support of the classic linguistic intuition of sparser temporal resolution in representing tones, Tian, Zhou, Chu, and Chang (2004)’s automatic tonal recognition study of Mandarin previously showed that recognition with just 4 samples/tone can outperform recognition with 1 sample/10 ms, concluding that “detailed information is useless for tone discrimination” (Tian et al., 2004, I-107).

Other Mandarin and Cantonese tonal recognizers have used a simple time warping-like (time normalization) sampling scheme of just 3–5 f0 averages or frame values over uniformly divided subsegments of (part of) the syllable (Peng & Wang, 2005; Qian, Lee, & Soong, 2007; Wang & Levow, 2008; Zhou, Zhang, Lee, & Xu, 2008). For the synthesis of natural-sounding Cantonese tones, Li and Lee (2007, 2008) argued that one or two linear movements per tone sufficed. These uses of just a few samples per syllable in feature extraction for tonal recognition are striking, in the context of the dominance of sampling rates typically an order of magnitude higher in automatic speech recognition. Still, the computational literature on tone recognition has not settled on the fineness

[3] The shape of a formant or f0 trajectory may also be parameterized in terms of a family of functions (Andruski & Costello, 2004; Hermes et al., 2006; Hirst & Espesser, 1993; Kochanski, Grabe, Coleman, & Rosner, 2005; Li & Chen, 2016; Prom-on, Xu, & Thipakorn, 2009; Shih & Lu, 2015; Taylor, 2000), such as quadratic polynomials, but the issue of temporal resolution remains. For instance, the more finely one wishes to capture the detailed shape of the contour, the higher the degree of polynomial needed.

of sampling resolution to use in tonal feature extraction from the speech signal.

But recent work on tonal production, perception, and processing has raised questions about the assumption of the sufficiency of coarser sampling for tones. Barnes, Veilleux, Brugos, and Shattuck-Hufnagel (2012) argued that details of contour shape such as convexity and concavity define English intonational pitch accent contrasts. Remijsen (2013) and Remijsen and Ayoker (2014)discovered contrastive early/late falls in Dinka and Shilluk, and DiCanio, Amith, and Garća (2014) found contrastive early/late rises in Yoloxóchitl Mixtec. In Dinka and Shilluk, a difference of a mere 40–64 ms in the onset of an f0 fall signals the lexical contrast between an early fall and a late fall. Remijsen (2013) also provided evidence that Dinka speakers had no trouble discriminating between the early and late falls in perception. Moreover, Chandrasekaran, Krishnan, and Gandour (2007) and Krishnan, Gandour, Bidelman, and Swaminathan (2009) found in electroencephalographic studies that Chinese speakers showed a language experience advantage relative to English speakers in processing rising f0 contours only when their curvature matched the concave shape of the rising tone in Chinese, and not when they were approximations of one or two linear segments or convex. Krishnan, Gandour, and Ananthakrishnana (2014, 2015, 2017), i.a. also found evidence suggesting that neural activity in the auditory cortex in Mandarin speakers is sensitive to changes in acceleration rates of rising and falling in the f0 contour, as well as the location of the turning point. In sum, there is a growing body of evidence that human tonal perception involves temporal resolution fine enough to capture details of the f0 contour shape.

### 1.2. Goals of current study

While the evidence presented in the previous section indirectly bears on the fineness of temporal resolution in tonal perception, there is little work that explicitly tests the effect of temporal resolution on tone perception. Our study does just that. Perceptual studies that present listeners with tones resynthesized with particular restrictive parameterizations, e.g. as polynomials of a certain degree, can provide at least a preliminary indication of whether the restrictive parameterization chosen could be a close approximation to that in human perception by collecting similarity judgments between the resyntheses and original stimuli (Hermes et al., 2006; Li & Lee, 2007). Here, we backed off from imposing a hypothesized restrictive parameterization on the listener and focused on the more general issue of assessing the effect of reducing the temporal resolution of the speech signal on tonal perception. We manipulated which timepoints in the stimuli the listeners had the opportunity to hear by intermittently deleting the recorded speech signal and replacing it with white noise, in the tradition of phoneme restoration (Bashford, Riener, & Warren, 1992; Miller & Licklider, 1950; Samuel, 1996; Warren, 1970) and "silent center" (Strange et al., 1983) perceptual experiments.

Gottfried and Suiter (1997), Lee (2009)and Lee, Tao, and Bond (2008, 2009) previously performed "silent center" Mandarin tonal perception experiments, which could be construed as manipulations of temporal resolution in the speech signal. Lee (2009) and Lee et al. (2008, 2009) built on Gottfried and Suiter (1997)'s small-scale study and included comparisons of tonal identification accuracy for listeners under time pressure between intact tones, "silent center" (Strange et al., 1983) tones with only initial and final regions available and the speech material in between silenced ("2 samples" over the tone; the first 6 and final 8 pitch periods), and tones with only the initial region available ("1 sample" over the tone; the first 6 pitch periods). Key results from these studies are that: (a) tonal identification accuracy did decrease as a function of the amount of input available to the listener, but remained high and well above chance (25%)—mostly 80%–95% accuracy—regardless of whether the stimuli were single speaker or multi-speaker, whether presented in isolation or with preceding context, and whether the preceding context was cross-spliced in from another recording or not; (b) providing preceding context significantly facilitated tonal identification for the onset-only and silent-center conditions, relative to providing no context; (c) reducing the amount of speech signal available to the listener affected different tones differently.

Our study builds on this work. First, we chose to perform the study in Cantonese avoid the ceiling effects that occurred with Mandarin. Tonal identification in Cantonese presents a more challenging tonal identification task than in Mandarin. While Mandarin has four tones which all have very different f0 contour shapes, the tonal inventory of Cantonese includes three level tones (high level Tone 1, T55, ˥; mid level Tone 3, T33, ˧; low level Tone 6, T22, ˨), two rising tones (high rising Tone 2, T25, ˧˥; low rising Tone 5, T23, ˨˧), and a falling tone (Tone 4, T21, ˩˧), cf. Fig. 2 (Matthews & Yip, 1994).[4]

Second, we make a more explicit connection between what acoustic information is available in the signal and how listeners perceive tones as the signal is degraded. To do this, we accompanied the human perception experiment with a machine classification study. To make the human and machine results comparable, we designed our study to try to simulate the conditions of machine classification in the perception task for the human listeners. We provided an experimental context for tonal identification limited in a way to be similar to characteristics of feature extraction in automatic tonal recognition. We used tritone stimuli from connected speech, as most recent automatic tonal recognizers use acoustic feature extraction from a temporal window extending beyond a single tone to its neighbors (Levow, 2005; Qian et al., 2007; Zhang & Hirose, 2000), and we used stimuli from multiple speakers like in the speaker-independent tonal recognition tasks in Qian et al. (2007) and Peng and Wang (2005). We also resynthesized the syllable durations of the tritones to be fixed at their grand average to simulate the commonly employed preprocessing step of time normalization to the syllable. Our experimental manipulation of temporal resolution in the signal used interrupting noise to create a 5-step gradient of sampling resolution and make uniformly distributed "samples" or windows from the speech signal available to the listener—a very simple treatment designed to simulate the common use of uniform

[4] Descriptions vary slightly in the exact 5-value integers assigned to the tones, but the exact integers used here are not of importance since we use these designations purely as mnemonic names. Some descriptions also distinguish these tones from the shorter entering tones (high, mid, and low level) which occur in syllables with unreleased stop codas. Throughout the paper, we use 5-valued integer designations as mnemonic names for the tonal categories, e.g. T55 for ˥.

sampling in feature extraction for automatic tonal recognition.

Finally, our study connects how often the speech signal is sampled to *where* the speech signal is sampled. Khouw and Ciocca (2007) found that f0 information at the syllable offset was the most critical in acoustic and perceptual discrimination of Cantonese tones in isolation. To follow up on this, while Lee et al. (2008, 2009)'s had a single 1-sample condition sampling from the target syllable onset, we compare sampling from the syllable onset, midpoint, and offset in our machine classification experiments. Also, while Lee et al. (2008, 2009) only preceded the target syllable with other speech material, we include a syllable *following* the target syllable, in addition to one preceding the syllable, as in Gottfried and Suiter (1997). This following syllable provides a buffer for f0 information in the target syllable offset that may be shifted or spread onto the following syllable.

Based on the past work discussed in this section, we had three general hypotheses for our study:

1. *Limited effects of decreasing resolution*: like Lee et al. (2008, 2009) found for Mandarin, tonal identification accuracy by humans and machine will be well above chance for all temporal resolution conditions, with little detriment to identification accuracy overall as the sampling becomes coarser.
2. *Tone-specific disadvantage with coarser resolutions*: the brunt of the deterioration in tonal identification as resolution drops will come from confusability involving the two rises, T25 and T23, which participate in subtle contrasts in f0 contour shape similar to that of the contrastive falls/rises found in DiCanio et al. (2014), Remijsen and Ayoker (2014), and Remijsen (2013).
3. *Informativity of syllable offset*: based on Khouw and Ciocca (2007), the availability of acoustic information from syllable offsets will facilitate perception more than information from syllable onsets or midpoints.

In the rest of this paper, we describe the speech materials used in the perception experiment and procedures for the experiment and analysis (Section 2), present results from the perception experiment and machine classification task (Section 3), discuss these results (Section 4), and conclude in Section 5.

## 2. Materials and methods

### 2.1. Recordings

The stimuli were recorded by ten native Cantonese speakers, five of whose recordings were further processed for the rest of the study: these three males and two females were chosen to span a wide pitch range (see Appendix A), to provide a representative instance of the challenge of a multispeaker task. Four of the speakers were born and raised in Hong Kong and recorded in the phonetics lab sound-attenuated booth at the City University of Hong Kong. One was born and raised in Macau and recorded in the phonetics lab sound-attenuated booth at University of California, Los Angeles. They were recruited from the local university student population and received cash compensation. All speakers were recorded using a Shure SM10A-CN headworn mic. For the speaker at UCLA, the signal was run through an XAudioBox preamplifier and A-D device to a computer at 22,050 Hz/16 bits with PCQuirerX (Scicon R&D, Inc.). The speakers in Hong Kong were recorded at 44.1 kHz/16 bits with a TASCAM HD-R1 digital recorder.

The stimuli were created from the tritone ⟨*waia*˥, {*waia*˦, a˧, a˥, a˩, a˨˩, a˧˥}, *mata*˥⟩ (*wai*$^{33}$ *wai mat*$^{3}$) extracted from sentences of the form: *lei*$^{25/35}$ *yiu*$^{33}$ *wai*$^{33}$ *wai mat*$^{3}$ *deng/geng*$^{33}$ '*you want Wai-Wai to clean the lamp/mirror*' with the target, the second /wai/, ranging over all six Cantonese tones. They were part of a larger study on contextual tonal variability. The lexical meanings of the orthographic characters we associated with tones T55, T25, T33, T21, T23, and T22 were, respectively, 'power', 'appoint', 'fear', 'surround', 'great', and 'stomach', and speakers were asked to treat /wai wai/ as a (nonce) proper name. The orthographic characters were chosen to be the most familiar ones for each tone by a native speaker. Each speaker actually recorded 5 fluent repetitions of sentences containing all 36 bitone combinations over /wai wai/ (with the sentences not used as stimuli for the perception experiment serving as fillers), from which we chose the last three repetitions of each Tone 33-Tone X-Tone 3 tritone for the stimuli set for a total of 90 tritones, 18 from each speaker, 3 distinct repetitions per speaker per tritone.[5] We held the initial and final syllable of the tritone constant as mid-level tones so that the listener's task was limited to the identification of a single tone, since a pilot experiment allowing variation in more than one tone was very confusing to participants. A Cantonese native speaker trained in linguistics and phonetics checked that none of the speakers had tonal mergers and that the speakers uttered the tones correctly. No speakers produced Tone T55 with a 53 high fall contour, a variant more common in the past.

### 2.2. Resynthesis

All stimuli were resampled to 22 kHz; tritones were extracted using a rectangular window, and RMS amplitude was rescaled to 75 dB (relative to the auditory threshold) in Praat (Boersma & Weenink, 2010). All syllables were resynthesized using PSOLA implemented in Praat to be have a target duration of 241 ms, the grand mean of the syllable durations, for a total duration of 740 ms for the tritone, to simulate time normalization to the syllable.[6] The manipulated condition, TEMPORAL RESOLUTION, was varied from the intact signal, to 7, 5, 3, and 2 uniformly spaced samples (time-slices or windows) of 30.41 ms each per syllable. The sample duration was well below the minimum 130 ms duration Greenberg and Zee (1977) found necessary for perception of a nonzero f0 velocity, "contouricity", in speech, and also on the same order of magnitude as the standard frame size in automated short-term analysis f0 detection (Hess, 1983, 343).

The temporal resolution manipulation involved intermittently deleting the recorded speech signal and replacing it with white noise low-pass-filtered at 5000 Hz that was 10 dB higher than the average signal amplitude, cf. Fig. 1. Similar stimuli manipulations are used in phonemic restoration studies, in which listeners perceive segmental speech sounds to be present in the presence of noise even if they are not (Bashford et al., 1992;

[5] In three cases, we chose another repetition than those listed above due to sound quality of the recording.

[6] The PSOLA algorithm resynthesis added about 18 ms over the target duration over the course of the tritone.

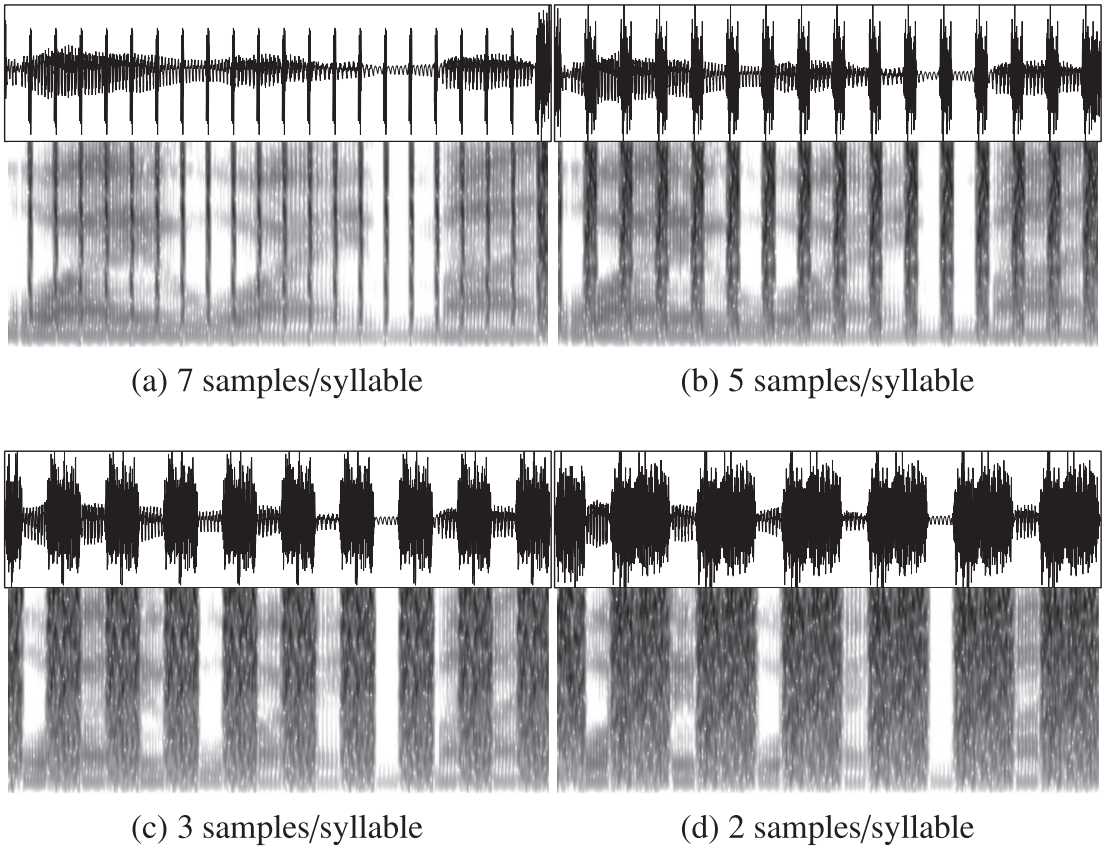


**Fig. 1.** Waveforms and spectrograms of a Cantonese tritone Tone T33 – Tone T21 – Tone 33 stimulus under different temporal resolution conditions from 7, 5, 3, and 2 samples/syllable. The intact condition is not shown.

Miller & Licklider, 1950; Samuel, 1996; Warren, 1970). We alternated the speech signal with louder noise rather than silence because the intelligibility of the speech is well-known to be poor when alternated with silent gaps; however, continuity of the speech percept can be maintained when the speech signal is alternated with a louder sound that is a potential masker of the fainter speech signal. This phenomenon is in fact the basis of phonemic restoration. Broadband noise has typically been used in segmental phoneme restoration experiments, and we chose to use white noise low-pass-filtered at 5000 Hz in particular because it has also been used in studying the continuity of tones through interrupting noise (Ciocca & Bregman, 1987). Additionally, we chose white noise to avoid providing any information that the listener might use in perceiving the interrupted speech, since Bashford, Warren, and Brown (1996) showed a boost in the intelligibility of speech interrupted by speech-modulated noise rather than white noise.

The noise was generated using the MLP Matlab toolbox (Grassi & Soranzo, 2009). Since the sample durations were fixed, the noise duration varied for different RESOLUTION conditions, but was fixed within a condition, ranging from 90 ms to 4 ms from the 2- to 7-sample condition, respectively, as shown in Fig. 1. The duration of noise intervals was measured in absolute time rather than the number of glottal pulses, cf. Lee et al. (2009), to simulate the constant frameshift used in feature extraction in automatic tone recognizers. The noise intervals included raised-cosine onset and offset ramps that were 10% of the duration of the noise interval to reduce audible spectral splatter (Hant & Alwan, 2003); the duration of the ramps was chosen to be relative to the duration of the noise interval since the noise interval duration varied between sampling resolution conditions. Half-duration noise intervals were used at the onset and offset of the tritone, with extra noise padding at the offset if needed to replace the entirety of the duration of the intact speech signal. Due to a programming error not detected until after the participants were tested, the last noise interval for the 2- and 3-sample stimuli was of full rather than half duration. For these two conditions, the final noise interval thus extended the stimulus duration beyond that of the stimuli for the other conditions. However, the same information from the speech signal was available to the listeners that would have been present without the added noise: the extended noise at the stimulus offset did not replace any speech information.

### 2.3. Acoustic feature extraction

Fundamental frequency (f0) timecourses were extracted for stimuli description and machine classification. The f0 values, shown in Fig. 2, were extracted using RAPT (Talkin, 1995), a commonly used f0 detection algorithm, used in Qian et al. (2007)'s Cantonese supratone tonal recognizer. Speaker-specific pitch floors and ceilings were set to the 1st and 99th quantiles minus or plus 30% of the range, respectively, a similar procedure to the pre-processing procedures in De Looze and Rauzy (2009) and Evanini and Lai (2010).[7] Otherwise,

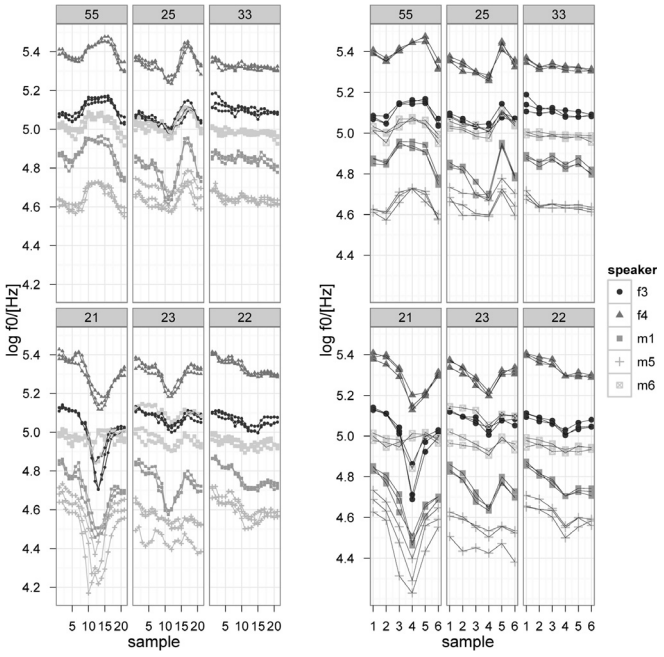


**Fig. 2.** Fundamental frequency (f0) contours extracted with RAPT using speaker-specific pitch floors and ceilings, showing the parameterization of f0 contours for the 7-sample and 2-sample conditions for computational modeling. Linear interpolation was used for replacing missing values and smoothing. Left panel: log-transformed f0 extracted with 10 ms frameshift and averaged over each of the 21 samples in the 7 samples/syllable condition. Right panel: log-transformed f0 values averaged over each of 6 samples in the 2 samples/syllable condition.

the default parameter settings, including a 10 ms frame shift were used. The first three and last frames were excluded because there were often large discontinuities between the estimated f0 for these frames and estimated f0 in the adjacent ones due to edge effects in the f0 detection algorithm, so there were a total of 67 f0 values, which were taken as the available f0 information in the intact condition. Unvoiced frames and frames with estimated f0 values resulting in large discontinuities were assigned f0 values using linear interpolation. To model the f0 information present in the degraded RESOLUTION conditions, the mean f0 was calculated over each unmasked region over frames falling within each of these regions. Thus, there were 6 f0 values estimated for each tritone in the 2-sample condition, one per unmasked region, and 9, 15, and 21 f0 values in the 3, 5, and 7-sample conditions, respectively. The f0 values were also log-transformed and then standardized as z-scores using speaker-specific means and standard deviations.

### 2.4. Participants

The participants were 39 native Cantonese speakers. There were 20 males (age 18.9 ± 1.8 years) and 19 females (age 21.9 ± 1.9 years). Participants were recruited from the local university student population at the City University of Hong Kong and at the University of California, Los Angeles and received cash compensation. All but three of the subjects (born/raised in Guangzhou and Shanwei, China) was born and/or raised in Hong Kong, China. Of the 10 participants tested in Los Angeles, all used Cantonese on a daily basis and had been in the United States for 3–8 years. No participants reported abnormal hearing.

### 2.5. Procedure

Participants were tested in sound-attenuated booths in the phonetics laboratories at the City University of Hong Kong and University of California, Los Angeles. The perception experiment was run in MATLAB using Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997). Stimuli were played from an Echo Indigo IO sound card on a laptop over studio monitor headphones at a standardized, comfortable volume, and the responses and reaction times of the subjects measured from the onset of the stimulus were recorded.[8] The inter-stimulus interval was 3 s.

Participants were told that the stimuli were extracted from sentences *lei*$^{25/35}$ *yiu*$^{33}$ *wai*$^{33}$ *wai mat*$^{3}$ *geng*$^{33}$ 'You want NAME to clean the mirror,' and they were given a sheet of paper with orthographic characters which showed what stimuli was being played, and what word they were to identify: *wai*$^{33}$ ___ *mat*$^{3}$. The stimuli were blocked by temporal resolution; block order was pseudorandomized to be roughly uniformly distributed over sampling resolution condition across participants to average over learning effects between blocks, and stimuli were randomized within blocks. Each block contained 90 stimuli, 18 from each of the 5 speakers, with 3 distinct repetitions per speaker per target tone. The task of the participants was to lexically identify the target syllable in each stimulus by a keyboard press of one of six keys labeled with the characters for the minimal tone set over *wai*. Participants were asked to respond as quickly and accurately as possible and told that they would be timed.

### 2.6. Data analysis

Statistical analysis was performed in R (R Core Team, 2014), and graphics were created using the ggplot2 package (Wickham, 2009). Tone identification accuracy was tested against at-chance levels (1/6) in each resolution condition by checking if 1/6 was contained in the 95% high posterior density interval of Bayesian estimates of the mean accuracy across tones (Kruschke, 2013). For the human perceptual data analysis, all listeners and items were included in the analyses. While all listeners performed at above chance levels in the intact condition overall, not all listeners performed above chance levels for each individual tone, and in addition, there were three items that were not identified at above chance levels.

#### 2.6.1. Machine classification

For insight into the human perception results, the acoustic separability of the different tones in the stimulus set was assessed using the acoustic feature extraction described in Section 2.3. These acoustic features were used in machine classification of the tone stimuli, using support vector machine classifiers (SVMs) (Burges, 1998; Cortes & Vapnik, 1995;

[7] The majority of the f0 values were in the mid range since each tritone stimulus consisted of two mid-level tones (33), yielding a center-heavy distribution of f0 values; thus, we could not use less extreme quantiles as in De Looze and Rauzy (2009) and Evanini and Lai (2010)because they resulted in severe compression of the estimated range.

[8] Reaction times were measured but not further reported here since the 2- and 3-sample conditions had longer stimuli than the other conditions and since no significant effects were found for temporal resolution.

Vapnik, 1995). SVMs are well-understood and widely used in machine learning and have been popular for automatic tonal recognition, e.g. Chen, Yang, and Liu (2014), Lam (2014), Levow (2005), Peng and Wang (2005), Peng, Zheng, and Wang (2004), Wang and Levow (2008) and Wang, Tang, Zhao, and Ji (2009). For an intuitive explanation of how SVMs work, see Appendix B.

Because the SVM algorithm involves calculating Euclidean distances in the parameter space, it is necessary to scale the data so that parameters with a greater range do not dominate how tones are classified, relative to parameters with a smaller range. Thus, the f0 data was log transformed and then z-score standardized, following Levow (2006, Section 2.3). (This preprocessing step was also similar to that of Peng & Wang (2005), which used a log-transformed 5-level normalization.) An SVM classifier was built for each of the 2-, 3-, 5-, and 7-sample conditions and 10 ms frameshift "intact" condition, as well as three 1-sample conditions, using either the first, second, or third sample of each syllable from the 3-sample condition, "1-sample-initial", "1-sample-medial" and "1-sample-final". These 1-sample conditions are included in the discussion only when explicitly mentioned.

Linear SVMs were implemented with LIBSVM (Chang & Lin, 2001) in R's e1071 package (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2012). The functions that linear SVMs use to classify data are linear combinations of the features, e.g., for a resolution condition with $n$ f0 samples in total over the three syllables, $\alpha_1 \cdot f0_1 + \alpha_2 \cdot f0_2 + \cdots + \alpha_n \cdot f0_n$, where, for all $i$, $\alpha_i$ is a constant. Thus, the absolute value of the feature weight can be used as a measure of a feature's relative importance for the classifier (Guyon & Elisseeff, 2003).

The 6-way Cantonese tone classification problem was decomposed as $\binom{6}{2} = 15$ binary classification sub-problems, e.g. T55 vs. T25, and the tone category receiving the most votes over all the sub-problems was selected as the classification decision. For each sampling resolution condition, the data was partitioned into 5 folds, one fold per speaker, for 5-fold cross-validation. Rotating across the folds, a single fold (18 tritones, 1 speaker) was used as training data, and the remaining four folds ($4 \times 18 = 72$ tritones, 4 speakers) were used as test data. All classification results, unless otherwise indicated, were averaged across the results from the 5 rotations, and standard error for classification accuracy was calculated from the variance of the accuracy over the 5 folds. Tonal classification accuracy was analyzed with logistic regression as described in Section 2.6.2.

#### 2.6.2. Logistic models for tonal identification responses

The probability of correctness of tonal identification was analyzed using mixed effects logistic regression implemented by the `lme4` package of Bates, Maechler, Bolker, and Walker (2014). Since both exploratory data analysis and logistic models interacting tone with resolution condition showed that the effect of RESOLUTION on tonal identification accuracy varied by tone (see Fig. 3), we analyzed separate models of identification accuracy for each of the six tones; models were separated by tone for ease of interpretability. The logistic models included the fixed effect of RESOLUTION; the inclusion of this fixed effect was justified by the experimental design since resolution was the critical variable of interest that was manipulated (Barr, Levy, Scheepers, & Tily, 2013). Contrasts for RESOLUTION were specified in two ways: (1) as treatment contrasts between each degraded RESOLUTION condition and the intact condition, i.e. the intact condition was set as the reference level, and (2) as forward difference contrasts between each successive RESOLUTION condition, i.e. between the 3-sample and 2-sample condition, between the 5-sample and 3-sample condition, etc.[9] Logistic models were also used to analyze the effect of RESOLUTION on the probability of response of a tonal category for the least accurately identified tones and the tones they were most confused with: responses of T22 and T23 for T33, responses of T25 and T22 for T23, responses of T23 for T22, and responses of T23 for T25.

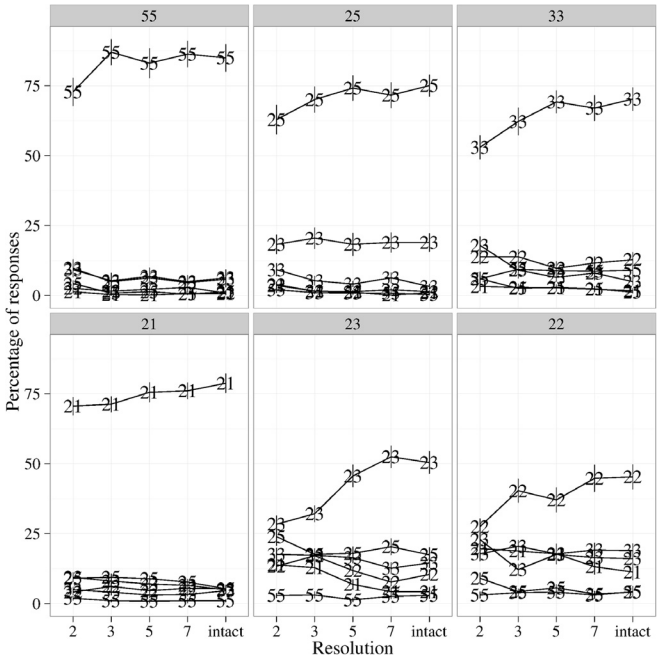


**Fig. 3.** Cantonese native listeners' tonal identification response frequencies for each of the six tones, conditioned on RESOLUTION. Accuracy for all tones except Tone T22 was significantly lower than in the intact condition only in the 2- and 3-sample conditions. T23 and T22 were identified with strikingly lower accuracy overall than the other tones were. Error bars show ±1SE over participants.

To avoid anticonservativity, the random effects structure was chosen to be the maximal random effects structure justified by the experimental design that led to convergence (Barr et al., 2013); this procedure resulted in the inclusion of random intercepts by (stimulus) speaker and listener, as well as random slopes for RESOLUTION by listener except when the model did not converge with the random slope included. For SVM data, we included random intercepts by speaker and fold, as well as random slopes for RESOLUTION by speaker when the model converged with them. Listed p-values for fixed-effects coefficients are from Wald z-statistics. Significance was determined at an alpha level of 0.05. (Full regression coefficient values and p-values are given in Supplementary materials, Section 5).

[9] Forward difference contrasts were defined following UCLA Statistical Consulting Group (2011). They compare the mean of the dependent variable for one level of the categorical variable to the mean of the dependent variable for the next level. See R code at https://github.com/krismyu/resolution for further details.

*2.6.3. Analysis of confusion matrices*

For visualizing confusion patterns in tonal responses, the confusion matrices from human perception and machine classification were also analyzed using the well-studied similarity choice model (Luce, 1963; Nosofsky, 1990), following Silbert (2014). This model allowed us to partition the underlying source of the tonal responses into the distinct contributions of similarity between stimuli and bias in responses. The calculated similarity matrices were used as input to non-metric 2-D multidimensional scaling (MDS) implemented using `smacof` (de Leeuw & Mair, 2009), and average linkage hierarchical clustering (James, Witten, Hastie, & Tibshirani, 2013, p. 390–396) using `hclust` (R Core Team, 2014). For the MDS solutions, stress was $5e^{-3}$ or below for all resolution conditions except for the intact condition, where stress was 0.01.

All R code used for analysis can be found at https://github.com/krismyu/resolution.

## 3. Results

This section describes results on the effects of temporal resolution on acoustic classification and perceptual identification of Cantonese tones. The discussion of these effects is broken into three sections addressing our hypotheses: (a) results that provide evidence for only limited effects of reducing temporal resolution on acoustic classification and perceptual identification (Section 3.1), (b) results that show that coarser resolution affects different tones differentially (Section 3.2), and (c) results that compare the informativity of the syllable offset vs. the midpoint and the onset (Section 3.3).

### *3.1. Limited effects of decreasing temporal resolution*

The evidence for only limited effects of decreasing temporal resolution comes from three results: (a) tonal identification was well above chance for humans and machine down to just 1–2 f0 samples/syllable (Section 3.1.1), (b) decreasing temporal resolution had quite limited effects on identification (Section 3.1.2), and (c) for both humans and machine, the multidimensional scaling and hierarchical clustering solutions were very similar across resolutions (Section 3.1.3).

*3.1.1. High identification accuracy down to 1–2 samples/syllable*

Tonal identification was well above chance (1/6 = 16.67%) for every resolution condition for humans and machines, even when humans had less than a quarter of the speech signal available with 2 samples/syllable, and when SVM input consisted of only 1 sample/syllable. For humans, it ranged from 67.46% (SE 2.91) in the intact condition to 60.51% (SE = 2.41) in the 3-sample condition and 52.54% (SE 2.90) in the 2-sample condition. With only 3 samples per syllable, human identification accuracy for T55 was 87%, 71% for T21, 70% for T25, and 62% for T33. SVM classification accuracy of z-scores ranged from 65.00% (SE 2.86) in the intact condition (with 67 f0 values per exemplar) to 63.61% (SE 3.16) for the 2-sample condition. Accuracy was still 62.78% (SE 2.96) even in the 1-sample-medial conditions: classification accuracy was 100% for T55, 72% for T25, and 67% for T21, and 58% for T33 and T22.

Fig. 3 displays the relative frequency of responses for each individual tone in human perception, conditioned on RESOLUTION. Each panel shows the distribution of responses for a given tone, indicated by the label at the top of the panel. The response tone is indicated by text labels at each data point, e.g. the identification accuracy for T55 can be read off from the series of points marked as “55” in the panel labeled “55”. Fig. 4 is like Fig. 3, but for SVM classification of z-score standardized log-transformed f0 values. These figures show that both humans and SVMs recognized T55 with the highest accuracy. However, while T55 accuracy for humans was around 85%, it was 100% across nearly all resolutions for the SVMs. Also like humans, the SVMs recognized T22 and T23 with around 30–50% accuracy, the lowest out of all tones, and T25 and 33 with 70–75% accuracy. SVM accuracy for T21 recognition was around 53%, though—much lower than human accuracy for T21, which was 70% and above.

The reader may wonder why accuracy in the intact condition was so low for both humans and SVMs. From Figs. 3 and 4, it is clear that for both humans and SVMs, the poor performance in the intact condition can be traced to errors on T22 and T23, and for SVMs, on T21, too. Identification accuracy was 43% for T22 and T23 in the intact condition for humans, but 77% for the other four tones. Moreover, identification accuracy for the other four tones dropped only to 73% on average in the 3-sample condition and 65% on average in the 2-sample condition.[10] SVM classification accuracy in the intact condition for T55, T25, and T33 was 81% on average, but 49% for the other three tones. Accuracies in the 2-sample condition were within 2% of the intact ones. We return to discussion of consistently low accuracies for particular tones in Section 4.1.

*3.1.2. Limited drops in accuracy with decreasing resolution*

Figs. 3 and 4 also show that decreasing sampling resolution resulted in only quite circumscribed drops in accuracy relative to accuracy with the full speech signal.[11] There were little effects on human tonal identification accuracy down to 5 samples/syllable, and pervasive detrimental effects only at 2 samples/syllable.

Below, we present results from logistic mixed effects models comparing the odds of correct tone identification in degraded RESOLUTION conditions to the odds in the intact condition. Results for humans are summarized in Fig. 5. Each panel shows (for the tone labeling the top of the panel) the ratio of the odds of correct tone identification for the degraded RESOLUTION conditions to the odds of correct tone identification for the “intact” condition, with estimated 95% confidence intervals; numeric values are given in Section 5.1 in the Supplementary materials. For cases where the confidence interval includes 1, there was no significant difference between the odds of correct tone identification and the odds in the “intact” condition.

For T55, only the 2-sample condition showed a significant difference in the probability of correct tone identification. For T25, T33, T21, and T23, both the 2-sample and 3-sample con-

[10] It is important that not only T33 but also three other tones showed relatively high accuracy in the intact condition and lack of sensitivity to sampling resolution, since it’s possible that the occurrence of two consecutive *wai*$^{33}$ syllables for the T33 stimuli may have induced repetition-related effects in perception.

[11] Confusion matrices with numerical values are given in the Supplementary Materials in Section 2.

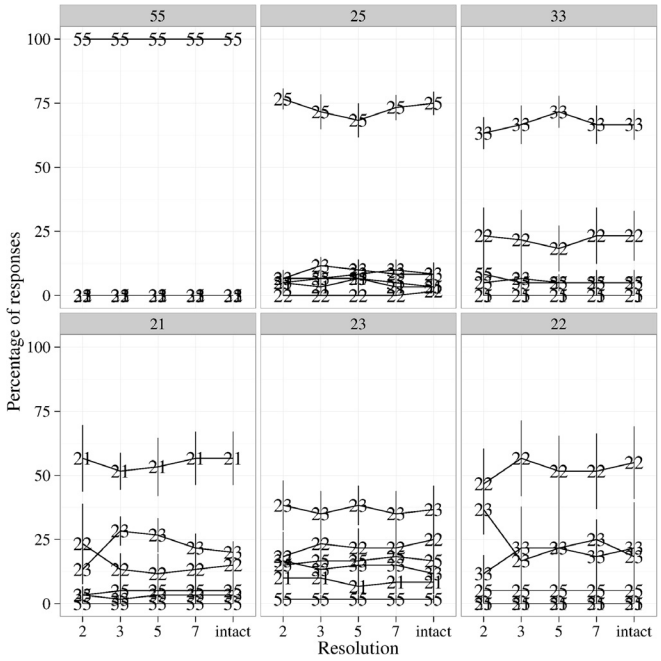


**Fig. 4.** Support vector machine classification response frequencies for each of the six tones, conditioned on RESOLUTION. Like in human perception, T55 was recognized with highest accuracy and T22 and T23 with the lowest accuracy. Unlike in human perception, T21 was also identified with relatively low accuracy. Error bars show ±1SE over folds.

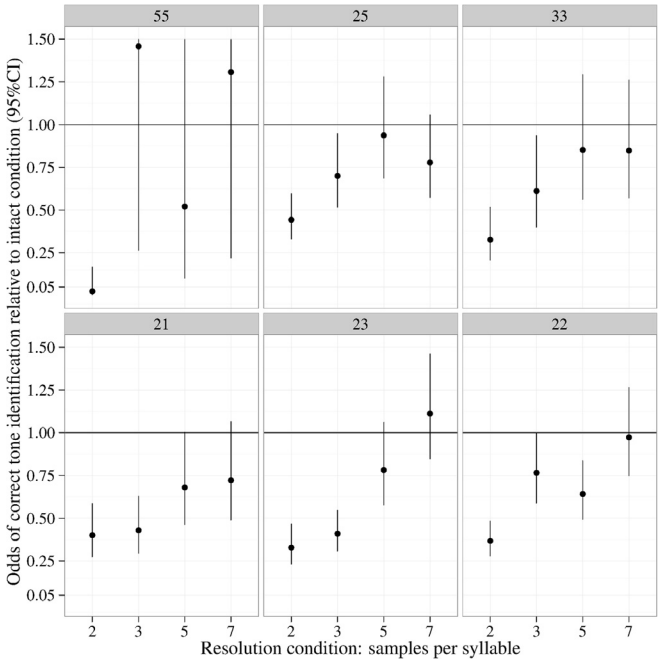


**Fig. 5.** Ratio of odds of correct tone identification by humans for RESOLUTION conditions to odds of correct tone identification by humans for the "intact" condition, with estimated 95% confidence intervals. Confidence intervals estimated using standard errors of fixed effects.

ditions showed a significant difference. For T22, 2-, 3-, and 5-sample conditions showed a significant difference. Thus, for all tones except T22, there were significant drops in tonal identification accuracy in the degraded resolution conditions only for the 2- and 3-sample conditions. In fact, for T55, T25, T33, and T22, the 95% confidence interval for the odds of correct tone identification relative to in the intact condition was 0.94 or above, even in the 3-sample condition. Temporal resolution therefore had pervasive detrimental effects on human tonal identification accuracy only in the 2-sample condition.

The effect of decreasing resolution in SVM classification was even weaker than in human perception (See Section 5 in Supplementary materials for SVM logistic regression results). In logistic models for SVM classification, we did not include T55 since it was recognized with near perfect accuracy across resolutions. We found significant effects for RESOLUTION only for T25: for z-scores, there was a significant decrease in accuracy for T25 in the 5-sample condition ($\beta = -0.53$ (SE 0.23), $z = -2.28$, $p = 0.02$). Between successive resolution conditions, only T22 showed any significant difference in odds of correct tone identification, and this drop in odds occurred between the 3- and 2-sample conditions.

### 3.1.3. Common acoustic and perceptual spaces across resolutions

More evidence for a limited effect of RESOLUTION on tonal identification comes from visualizations of the perceptual and acoustic spaces underlying tonal identification. Visualizations of the effects of RESOLUTION on the similarity matrices derived from the human responses are summarized in an overlay plot of 2-D MDS solutions for the different resolutions (Fig. 6) and in dendrograms from average linkage hierarchical clustering (Fig. 7). The gross distribution of tones in 2-D spaces derived from non-metric multidimensional scaling of human confusion data changed very little down to the 3-sample condition. T55 was quite isolated at the periphery of the space for all resolutions. The tones T55, T25 and T21 also consistently appeared rather aloof at the periphery of the MDS space across resolutions, while T33, T23, and T22 crowded together in the center of the space.

This same periphery-center distinction was mirrored in the dendrograms in Fig. 7. The height at which tone categories fused in the dendrogram indicates how similar they were: tone categories that fused near the top were very dissimilar from one another, while tone categories that fused near the bottom were very similar to one another. Thus, since T55 fused at the top of every dendrogram in Fig. 7, it was the tone that was the most perceptually dissimilar from all the other tones in every RESOLUTION condition. Tones T25 and T21 also fused near the top of every dendrogram, while the highly mutually confusable T33, T23, and T22 tones fused at the bottom of every dendrogram. The same was true for every dendrogram for SVM classification. (See Section 3 in Supplementary materialsfor SVM z-score dendrograms.)

The only major shift in the perceptual space down to the 3-sample condition was that T25 rose out of a bottom cluster with T23 from the 7- to the 5-sample dendrograms (Fig. 7, left). This was because the accuracy of correct T23 identification between those conditions dropped by 6.8%, a significant decrease in odds of correct identification ($\beta = -0.35$ (SE 0.15), $z = -2.30$, $p = 0.021$). Between the 7- and 5-sample conditions, T23, T22, and T33 retracted downwards into the interior of the perceptual space, away from T25, as shown in Fig. 6.

The distribution and location of tones in 2-D multidimensional scaling solutions for SVM confusion data was almost insensitive to RESOLUTION down to 1 sample, especially if only

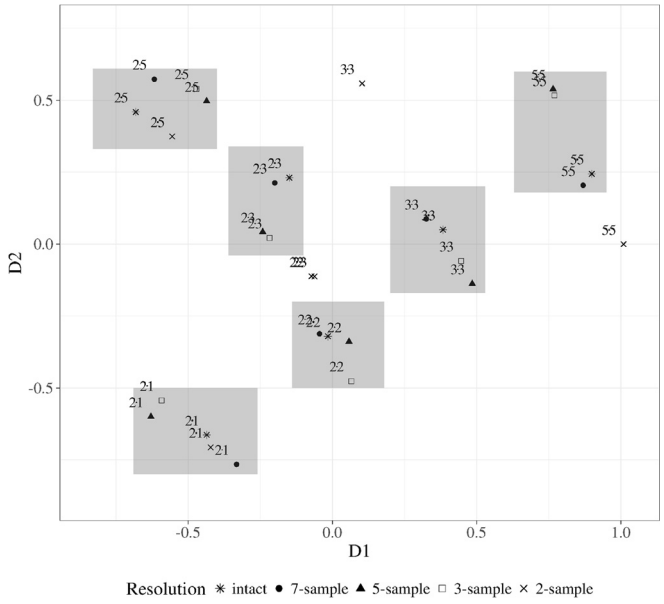


**Fig. 6.** 2-D multidimensional scaling solutions for similarity matrices derived from human tonal perception responses for each RESOLUTION condition. Regions enclosing a cluster of identical tones are highlighted with grey boxes, and text labels indicating tone labels corresponding to plotted points are offset upwards and to the left of the points. The 2-sample MDS space is quite different from all others, as evidenced by four of the six 2-sample points being located outside the grey boxes.

the 1-sample-medial condition was included; see Supplementary material Section 4. Points for any particular tone were tightly clustered together across resolutions. Like the human MDS space, the SVM space had T33, T23, and T22 right next to one another, and T55 isolated on the periphery. The one major difference from the human space was that the SVM space has T21 and T25 on the same side of the periphery next to one another. This indicates that T21 and T25 were much more similar to one another for the SVMs than for humans.

### 3.2. Effects of decreasing sampling resolution for particular tones

Close examination of confusion patterns, SVM feature weights, and f0 contours suggests that loss of detailed f0 temporal alignment and contour shape information with decreased resolution was a major cause of decreasing accuracy in tonal identification for identifying a subset of the tones. We describe three examples of evidence for this below: (a) the collapse of different f0 minima alignments between rising contours under coarser resolution (Section 3.2.1), (b) the role of loss of information about contour shape in the drop in T22 identification accuracy between the 3- and 2-sample conditions (Section 3.2.2), and (c) the consequences of undersampling of the f0 extremum for T23 vs. T25 and T21 vs. T23 classification (Section 3.2.3).

#### 3.2.1. Collapse of differing alignment of f0 minima in rising contours

The circled minima of the f0 contours of T25, T23, and T22 in Fig. 8 highlight the effect of decreasing resolution on the discriminability of rising f0 contours. As resolution decreased, the differences in the timing of the turning point in the f0 contours vanished. The f0 minimum of T25 was shifted later, to coincide with the minima of T23 and T22; the broad f0 valley of T22 was shifted earlier and sharpened, reducing T22's distinctness from T23; the depth of the f0 minimum for T23 was raised, reducing the distinctness between the T23 minimum and that of the other two tones. In the 2-sample condition, the circled points for all three tones appear in a vertical line at sample number 4: there's no more distinction between the alignment of the f0 minima of the tones. The collapse of f0 minima alignment and shape resulted in increased confusability of T22 and T23 and of T23 and T25—this is discussed in Sections 3.2.2 and 3.2.3.

#### 3.2.2. Contour shape and confusability of T22 and T23

Sampling f0 from only syllable onsets and offsets (the 2-sample condition) significantly reduced tonal identification accuracy from sampling medially as well as at onset and offset (the 3-sample condition). In the 3-sample condition, only T21 and T23 saw identification accuracy drop relative to conditions with finer sampling for humans (see Fig. 5). But sampling resolution had pervasive detrimental effects for every tone in human tonal identification only in the 2-sample condition.

Confusion patterns in human perception also shifted greatly from the 3- to 2-sample condition. This shift is most striking in

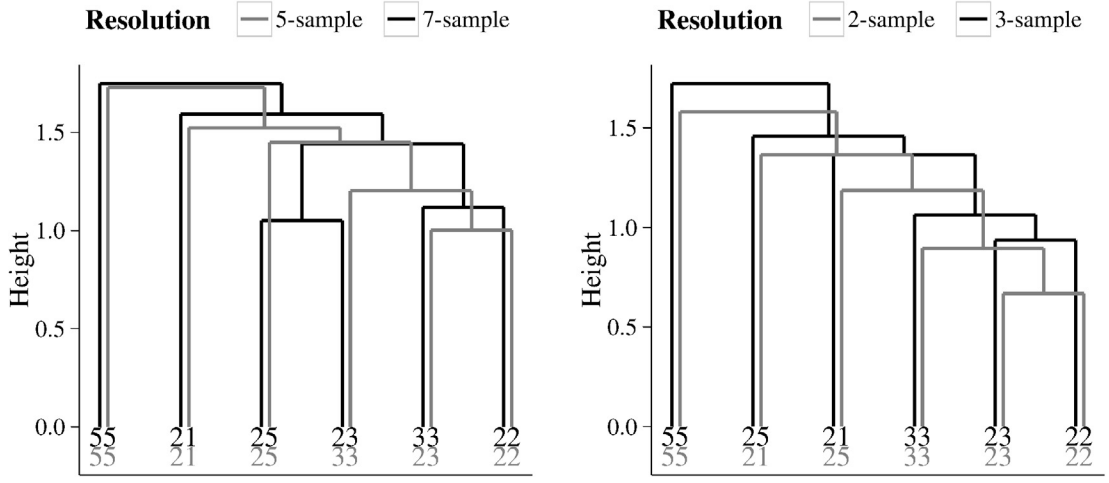


**Fig. 7.** Dendrograms from average linkage hierarchical clustering of similarity matrices. Similarity matrices were calculated from the similarity choice model applied to human perception data.

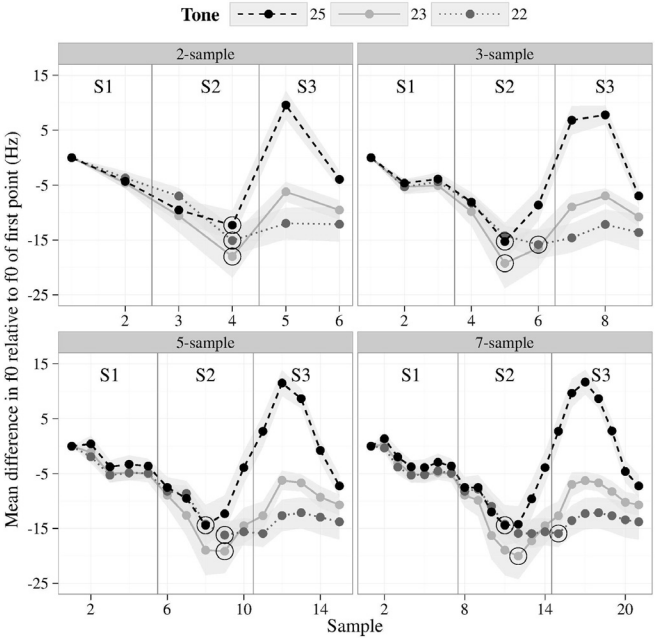


**Fig. 8.** Effects of sampling resolution on resolving turning points for T25, T23, and T22 stimuli. Each panel displays the f0 information present for the RESOLUTION indicated at the top of the panel, with linear interpolation between points. For each individual stimulus, the f0 value of the first point was subtracted from each f0 point. Each tone's f0 contour is averaged over all stimuli for the tone, and ribbons show ±1SE. A circled point indicates the global minimum in the aggregate contour.

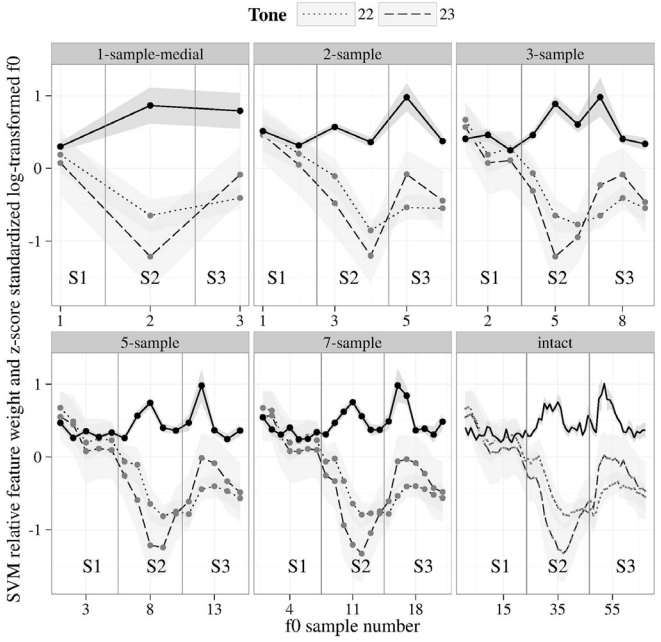


**Fig. 9.** Averaged z-score standardized log-transformed f0 contours (in grey) for T23 and T22 with time-aligned SVM feature weight magnitudes (in black) for different resolution conditions. Ribbons show ±1SE.

the 2-sample MDS space (Fig. 6), where the periphery-center divide in tonal distribution was destroyed from the 3- to 2-sample condition. T22 and T23 both shifted close to the origin (0, 0), nearly coinciding in the center of the space, but T33 shifted out into the periphery, towards T25. The high confusability between T22 and T23 was also reflected in the 2-sample dendrogram (Fig. 7, right), where the height of fusion for T22 and T23 was by far the lowest. Moreover, the odds of correct identification of T22 significantly decreased from the 3- to 2-sample condition—for humans, from 40.34% to 27.52% ($\beta = -0.58$ (SE 0.21), $z = -2.84$, $p = 4.6e^{-3}$) and for SVMs, from 56.67% to 46.67% ($\beta = -2.9$ (SE 1.06), $z = -2.72$, $p = 6.5e^{-3}$). This was the only significant change for SVM in odds of correct tone identification between successive resolution conditions. At the same time, the odds of a T23 response to T22 significantly increased for humans by 10% ($\beta = -0.58$ (SE 0.21), $z = -2.84$, $p = 4.6e^{-3}$) and for SVMs by 20% ($\beta = 2.60$ (SE 1.05), $z = 2.48$, $p = 0.013$).

Why were T22 and T23 increasingly confusable as resolution dropped? Fig. 9 suggests that the source of the confusion was the collapse of distinctions in the alignment and shapes of the T23 and T22 valleys. The figure shows the average z-score standardized log-transformed f0 contours for T23 and T22 (in grey) under different resolution conditions, together with time-aligned timecourses of SVM feature weight magnitudes for T22 vs. T23 discrimination (in black). Peaks in this timecourse indicate regions of particular importance for discriminating between T22 and T23. Thus, the location of the peaks in the SVM feature weight timecourses show that there were two time intervals of particular importance in the acoustic discrimination of T22 and T23: the descent to the f0 valleys in syllable 2, and the f0 peaks in syllable 3. In the 2-sample condition (top center panel), the relative importance of f0 information from the descent to the valleys (samples 3 and 4) plummeted.

A comparison of f0 contours across resolutions shows that f0 information from the 2-sample condition missed the difference in slope of descent between the sharp fall of T23 and the flat basin of T22. The lack of this information may have been the source of the increased confusability of T22 for T23 between 3- and 2-sample conditions in humans and SVMs (see bottom right panels in Figs. 3 and 4).

### 3.2.3. Undersampling the T25 and T23 f0 peaks

For classification between T25 and T23 (see Fig. 10), it was the time interval where f0 peaks occurred that had the most important f0 information.[12] As the signal was degraded, the number of samples where T25 and T23 were at their f0 peaks dropped from many in the "intact" condition to 2 samples in the 3-sample condition and just 1 sample in the 2-sample condition. Correspondingly, the peak in SVM feature weight coincident with the f0 peaks sharpened from a plateau to a point as the signal was degraded. Having just a single sample to detect the differences in peak height between T25 and T23 in the 2-sample condition may have been behind the significant 5% increase in T25 responses for T23 stimuli from the 3- to 2-sample condition ($\beta = 0.44$ (SE 0.15), $z = -2.86$, $p = 4.3e^{-3}$). There were no significant differences in odds of a T25 response for T23 between any other successive resolution conditions. For SVMs, there were no significant differences in odds of a T25 response for T23 at all.

[12] Like in Fig. 9, there is also a secondary peak in the SVM feature weight timecourse in Fig. 10. The secondary peak precedes the rise in the middle of syllable 2 and is missing in the 2-sample condition. However, this is actually because of a missing valley in the SVM feature weight timecourse between the two peaks: the 2-sample condition misses any f0 samples between the valleys and peaks of the T23 and T25 contours. In that time interval, the T23 rise can be quite steep, "catching up" to the level of the T25 rise after having dropped to a lower valley than T25.

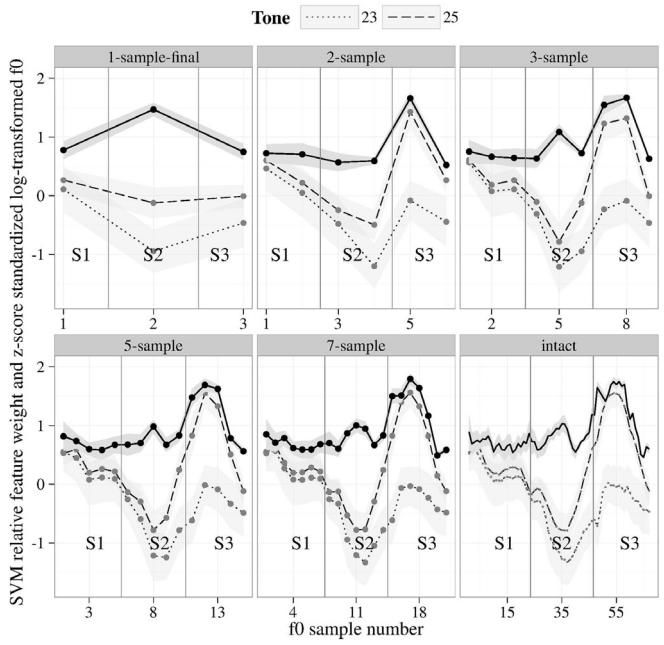


**Fig. 10.** Averaged z-score standardized log-transformed f0 contours (in grey) for T23 and T25 with time-aligned, scaled SVM feature weight magnitudes (in black) for different resolution conditions. The points were linearly interpolated. SVM feature weight magnitudes within a panel were scaled by a constant to make their relative differences easier to discern. Vertical lines within each panel delineate syllable boundaries. Ribbons show ±1SE.

### *3.3. Comparison of syllable onset, midpoint, and offset as locations for sampling*

Comparing tonal classification by machine under sampling just once per syllable at syllable onset, midpoint, and offset over the tritone (for a total of three samples over the entire stimulus), we found that: sampling at just the onset resulted in missing the valleys of T25 and T21 (Section 3.3.1); sampling at just the midpoint made rises and falls very confusable (Section 3.3.2), and sampling at just offset was severely detrimental to T25 identification (Section 3.3.3).

#### *3.3.1. Sampling at syllable onsets misses valleys*

Sampling at just the syllable onset resulted in poor tonal classification. While T55 identification accuracy was perfect for the other 1-sample conditions, it was only 70.00% (SE 5.65) when sampling only at syllable onset. This is because T55 became confusable with T25, since sampling at the onset results in missing the dip before the rise of T25. This can be seen in the f0 points for syllable 2 (S2) in the 3-sample panel in Fig. 8, which shows f0 contours for T25, T23, and T22 aggregated over all stimuli in the experiment by tonal category, for all degraded resolution conditions. (For easier interpretability, the f0 contours was standardized by subtracting off the first f0 value in the contour (the "anchor") from all other f0 values rather than shown as z-scores.) Fig. 8 shows that sampling just sample 4 (at syllable 2 onset) reveals no evidence of the f0 dip in syllable 2 for T25. Morever, since sampling at the onset results in missing the valley of T21, T21 was particularly confusable with T23 and T22 and classified correctly 36.67% (SE 11.96).

#### *3.3.2. Rises and falls: low f0 at syllable midpoint*

Sampling at just syllable midpoint resulted in shrinking the distinction between rises and falls. Rises and falls were not distinguished by L vs. H targets at the syllable midpoint. The f0 contours in Fig. 11 show that the T21 fall did not have a high target near syllable midpoint in syllable 2—it was high only at onset of the second syllable. Instead, the T21 fall had an f0 minimum between the syllable midpoint and offset in this syllable. Moreover, with decreasing sampling resolution, this f0 minimum appeared to shift to the right.

The location of the f0 minimum for T21 was also just where the f0 contours for T25, T23, and T22 achieved their f0 minima as well. This can be seen from Fig. 8. The f0 contours for these three tones descend to their f0 minima between the syllable 2 midpoint and offset, before beginning their ascents.

T21 was in fact most confusable with rising tones in human and SVM classification. In the human perception 2-sample condition, the frequency of responses of T23 and T25 to T21 was 9% for each rise, at least twice as much as responses for any other tone. Sampling only at the syllable midpoint, SVM identification of T23 was only 22%, with a 22% rate of T21 responses. In contrast, sampling at either the syllable onset or offset resulted in only 3–5% T21 responses. From the top left panel in Fig. 11, we can see that sampling at only syllable midpoint shrunk the difference in f0 heights at the f0 minima of T21 and T23 in syllable 2.

#### *3.3.3. Insufficiency of syllable offsets for identifying T25*

Sampling at just syllable offset resulted in shrinking the distinction between T25 and level tones. Sampling f0 from only syllable offsets was severely detrimental to identification of T25 in SVM classification. With just 1 sample per syllable, SVM classification accuracy for T25 was only 25.00%

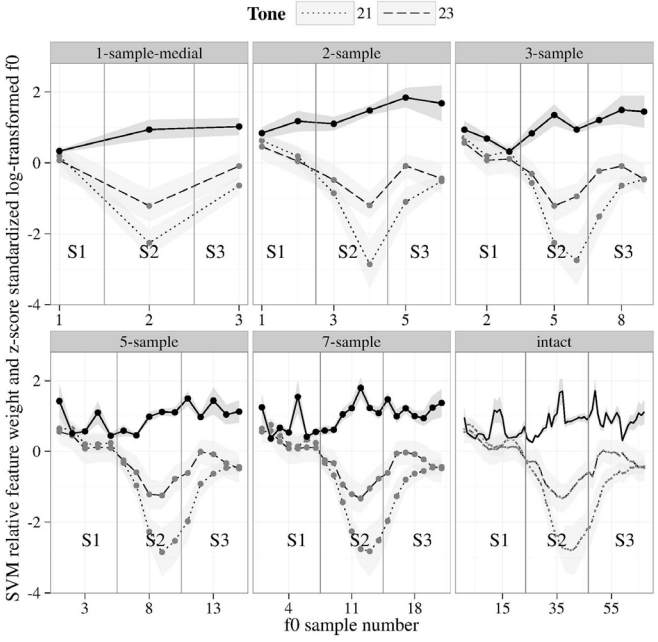


**Fig. 11.** Averaged z-score standardized log-transformed f0 contours (in grey) for T21 and T23 with time-aligned SVM feature weight magnitudes (in black) for different resolution conditions. Both tones had f0 minima between the syllable midpoint and offset.

(SE 6.45) at syllable offset, compared to 53.33% (SE 10.07) at syllable onset and 71.67% (SE 4.25) at syllable midpoint. In the 1-sample-offset condition, T25 was classified as the level tones T22 and T23 55% of the time. In comparison, T25 was mistaken for T22 and T23 only 15% of the time for the other 1-sample conditions.

The reason for the insufficiency of f0 information from the syllable offset is clear from Fig. 10. This is a timecourse plot of z-score standardized log-transformed f0 points overlaid with scaled absolute values of SVM feature weights for each f0 sample for T25 vs. T23 classification. The f0 contours are aggregated over all stimuli in the experiment by tonal category. Fig. 10 shows that the T23 and T25 rises both exhibited peak delay, with their peaks occurring in the syllable following the one that they were associated to (in syllable 3). Moreover, the f0 values most important in machine classification of these tones occurred at these peaks in the first half of the following syllable—not at syllable offset. When f0 information was sampled from just the offset (1-sample-final condition, top left panel), the ascent and descent of the T25 peak was missed almost entirely. This is why T25 was highly confusable with level tones in the 1-sample-final condition. The fidelity of the shape of the T23 contour with just syllable offsets was higher since f0 coming out of the f0 minimum was still relatively low at syllable offset in the 2nd syllable.

## 4. Discussion

This study provided evidence that decreasing temporal resolution in the speech signal had only limited effects on the acoustic classification and perceptual identification of tones. The evidence came from native speaker perceptual identification of Cantonese tones degraded by interrupting noise, as well as acoustic classification of the tones by support vector machines from fewer and fewer f0 data points in the input.

As we hypothesized, tonal identification accuracy by humans and machine was well above chance for all temporal resolution conditions:

1. Tonal identification accuracy above 70% was maintained for T55 and T25, even with just 3 samples/syllable for humans and 1 for SVMs, and accuracy for each tone in every resolution condition was well above chance.
2. Decreasing sampling resolution resulted in only limited decreases in identification accuracy—pervasive deterioration of identification accuracy occurred only at 2 samples/syllable for humans, and SVM classification accuracy with 1 sample/syllable from the syllable midpoint was indistinguishable from accuracy with 67 f0 values/syllable.
3. Acoustic and perceptual spaces for tonal identification computed from multidimensional scaling and hierarchical clustering remained very similar across resolution conditions, except for the 2-sample condition in human perception.

We also hypothesized that decreasing temporal resolution would differentially affect different tones, and that identification of the two rising tones T25 and T23 would be especially affected. This was indeed the case:

1. As resolution decreased, there was a collapse of differing alignment of f0 minima in the rising contours, making them increasingly confusable.
2. Dropping temporal resolution from 3 to 2 samples per syllable made T22 and T23 highly confusable for humans; SVM results suggest that this was because of loss in detail of contour shape at the f0 valleys.
3. SVM results suggest that increased confusability between T23 and T25 in humans from the 3- to 2-sample condition was due to loss of information about the shape of the f0 peaks.

Finally, we hypothesized that under sampling one f0 value per syllable, sampling at the syllable offset would facilitate tonal classification more than sampling at the syllable onset or midpoint. However, we found instead that sampling at any one of these points resulted in poor identification of some subset of tones. In particular, sampling at the syllable offset resulted in very poor accuracy in identification of T25 by SVMs because of peak delay which placed the peak in the first half of the syllable following the syllable T25 was associated to.

In the remainder of this section, we discuss results for each of the hypotheses: limited effects of decreasing temporal resolution in Section 4.1, as well as differential effects of decreasing resolution for particular tones and the effect of varying the location of sampling in Section 4.2. We then close the section by discussing the role of time in tonal phonetic spaces—temporal resolution and beyond—in Section 4.3.

### 4.1. Limited effects of decreasing temporal resolution

Our positive results for only limited effects of decreasing temporal resolution support the hypothesis of Chao (1930) and others—namely, that no more than a few f0 samples per syllable are needed to provide distinctive features for tonal categories. The programming error in the stimuli noted earlier that introduced longer noise intervals at stimulus offset in the 2- and 3-sample conditions (27 and 7 ms, respectively) does not weaken the result that decreasing resolution has a limited impact on tonal identification. The effect of the error could have only been in the direction of decreasing accuracy at the two lowest RESOLUTION conditions due to interference and/or memory effects.

In addition, the significantly lower accuracy in the 2-sample condition compared to that in the intact condition may have been due in part to a lack of perceptual continuity caused by the long duration of the interrupting noise intervals for our particular experiment design. (Recall that manipulation of the sampling resolution involved an increase of the duration of the noise interval as sampling resolution decreased.) In support of this conjecture, Dannenbring (1976) showed that in nonspeech, for pure tones of 250 ms in duration interrupted by white noise, the mean continuity threshold between perceived continuity and discontinuity due to the interrupting noise was around 80 and 100 ms for steady state tones and tone glides, respectively. This indicates that the 90-ms noise interval duration in the 2-sample condition may have been close to the auditory threshold for perceiving continuity in our stimuli, which were 241 ms in duration.

A potentially more serious challenge to the result that decreasing temporal resolution had only limited effects on tone identification is that the relative insensitivity of tonal identification accuracy to sampling resolution may have been due to a “floor effect”, i.e., already poor accuracy in the intact condition, which then left little room for further deterioration

with decreasing sampling resolution. As discussed in Section 3.1.1, poor overall performance by humans and SVMs in the intact condition was due to errors on T22 and T23, and for SVMs, errors on T21 as well.

We conjecture that the low accuracy for T22 and T23 in human perception was due in part to tonal mergers in some of the listeners[13] (Bauer, Kwan-hin, & Pak-man, 2003; Mok & Wong, 2010a, 2010b). However, there is also evidence that these tones were particularly confusable with other tones in the low pitch range when coarticulated with neighboring mid-level tones in the Tone T33 – Tone X – Tone 3 experimental stimuli. The strongest piece of evidence is that the SVMs performed just as poorly as the humans for T22 and T23. The SVM classification assumed that all six tonal classes were equiprobable and received equal numbers of exemplars for each tone class from each speaker: there is no sense in which there were tonal mergers for the SVMs. This implies that poor accuracy for T22 and T23 in SVM classification must have been a property of their acoustics.

If T23 and T22 were confusable when flanked by mid-level tones, then what about Tone 21? We hypothesize that Tone T21 was identified with high accuracy unlike its close neighbors T22 and T23 not only because it occupied a lower part of the pitch range than them, but also because of creaky voice quality cues, since Yu and Lam (2011, 2014) showed that the presence of creaky voice cues can boost T21 identification accuracy in Cantonese tone perception. In support of this hypothesis, identification accuracy for T21 was relatively high for humans, but low in machine classification. Also, the key difference between the MDS spaces for humans and SVMs was that T21 and T25 were much closer in the SVM acoustic space.

Most of the unvoiced frames in the RAPT f0 extraction came in T21 stimuli, since Tone T21 realization frequently had nonmodal phonation, low amplitude, and even intervals of silence. Acoustic feature extraction did not capture this, and while there are more sophisticated rules for estimating the pitch percept in the presence of nonmodal phonation for the human listeners than the simple linear interpolation used over voiceless frames that we used, cf. the aberrant pitch contours for T21 in Fig. 2, the poor accuracy for T21 identification by machine nevertheless suggests that a parameterization of the speech signal that references voice quality, beyond f0, is needed for both higher classification accuracy of T21 by machine, and for modeling what humans are doing.

Since the machine classifiers only had access to f0 information, while listeners had access to whatever information they were able to extract from the speech signal, the divergence between human and machine performance in classification accuracy of T21 also suggests a possible perceptual explanation for the introduction of creaky voice cues for T21. T21 was the only tone among the tones in the low range (T22, T23, T21) for which we did not observe a “floor effect” in human perception: identification accuracy for T21 in the intact condition was 79%, compared to 45–50% for T22 and T23. As pointed out by a reviewer, cases where cue redundancy is insufficient for coarse temporal resolution could lead to language change or the introduction of other cues, such as the creaky voice cues for T21.

Finally, a potential concern for the relevance of the SVM results for understanding human perception of tone is that providing z-scores as input gave the machines an unreasonable amount of information since the calculation of z-scores assumes the ability to parse speaker identity from the signal, as well as some knowledge about each speaker’s pitch range. However, it could be the case that z-scores might be a rough proxy for some constellation of information for calculating pitch range available to listeners, such as absolute f0 and aspects of voice quality, as well as previous experience with many speakers (Bishop & Keating, 2012). Moreover, classification accuracy for anchored f0 values using information only internal to each stimulus for pitch range information (see Fig. 8), was just as high, ranging from 65.28% (SD 10.44) in the intact condition to 63.61% (SD 9.99).

### 4.2. Effects of decreasing resolution for rising tones and the effect of sample location

Our poor SVM classification results under sampling just once per syllable bear on previous work characterizing tone using 1 or 2 f0 samples from particular locations in the syllable. First, the poor SVM classification results for T25 in tritones (i.e. uttered in connected speech) with 1 f0 sample per syllable offset provide a contrast with Khouw and Ciocca (2007)’s results that f0 change over the 6th and 7th out of 8 subsyllabic segments accounted for about 70% of the variance in a Cantonese tonal identification perception experiment of isolated monosyllables. While we did not use f0 change in the acoustic feature set, the bottom center panel in Fig. 10 suggests that f0 change at syllable offset might not have fared any better than f0 for discriminating T25 and T23, since both rises have similar slopes at syllable offset. Our results on the poor separability of T25 at syllable offset are also interesting in the context of a body of previous work on tonal coarticulation in Cantonese (Li, Lee, & Qian, 2002, 2004; Wong, 2006a) and other (South) east asian languages (Mandarin: (Xu, 1997), Thai: (Gandour, Potisuk, Dechonkit, & Ponglorpisit, 1992), Vietnamese (Han & Kim, 1974)). This literature has reported that rightward (carryover) coarticulation is stronger than leftward (anticipatory) coarticulation, so that tones in connected speech might be maximally separated near the offset of the syllable. It seems that if there is significant peak delay, as was the case is here, the region of maximal separation can get pushed into the first part of the following syllable. This could happen quite frequently, as there is likely to be peak delay under a wide span of speech rates: rising contours in at least Mandarin and Cantonese begin rising close to the syllable offset, even under slower speech rates (Wong, 2006b; Xu, 1998, 2001).

Second, while rises and falls in connected speech might be distinguished by L vs. H targets at syllable midpoint in Thai (Zsiga & Nitisaroj, 2007), this does not appear to be the case in Cantonese. At least in the T33 - X - T33 context of our experimental stimuli, both rises and falls had f0 minima near the target (X) syllable midpoint (Figs. 8 and 11) – in fact, the T21 fall’s f0 minimum was rather close to syllable offset. While both T25 and T23 rises and the T21 fall could be described as having L targets between syllable midpoint and offset, the f0 minima for

[13] For a discussion of potential tonal mergers in listeners and other possible reasons for low accuracy for T22 and T23, see Supplementary materials Section 6.

the three tones are, on average, at rather different heights. Thus, although rises and falls are not distinguished by different phonological targets, i.e., H vs. L, they are distinguished by different phonetic targets.

Third, although Barry and Blamey, 2004 proposed a simple 2-D space for Cantonese tone with the dimensions of f0 at syllable onset and offset, the high increase in confusability between T22 and T23 from 3- to 2-samples/syllable for both humans and SVMs suggests that something else not captured by f0 at syllable onset and offset also plays a role in Cantonese tonal perception. The SVM results (Fig. 9) show that the relevant information lost may be the shape of the valleys in the f0 contours. The use of creaky voice cues in Cantonese T21 perception also implies that the perceptual space for (native) Cantonese tone perception goes beyond a 2-D space of f0 samples, see Yu and Lam (2011, 2014).

### *4.3. Temporal resolution and beyond: the role of time in phonetic spaces*

If we were to include acoustic parameters relevant for the perception of creaky voice in the definition of the Cantonese tonal phonetic space, these parameters would also be subject to the issues on temporal resolution of f0 contours we have raised in this paper. These issues apply to *any* properties measured over the course of time[14]—including non-f0 based parameters that have been shown to be used in tonal perception, such as amplitude (Fu & Zeng, 2000; Fu, Zeng, Shannon, & Soli, 1998; Whalen & Xu, 1992) and other spectral measures, such as amplitude differences between harmonics (e.g. H1-H2, H2-H4, H1-A1) and cepstral peak prominence (Andruski, 2006; Andruski & Ratliff, 2000; Garellek, Keating, Esposito, & Kreiman, 2013; Kuang, 2013).

If, as this study suggests, particular details of contour shape and alignment can be important for accurate tonal classification, why not just sample as finely as feasible—for f0, for amplitude, for spectral balance measures, any acoustic parameter that might be relevant? Limited computing resources isn't truly the barrier to this: after all, it's the astonishing gains in modern computing power that has enabled neural networks to increasingly dominate the artificial intelligence world in recent years. The real issue is that having to sample everything as finely as we can to define a tonal phonetic space implies that we understand nothing about how tonal perception could work. To understand what humans are doing, we need to limit the number of parameters over which we define a phonetic space for scientific interpretability, or more generally, we need to understand how to weigh the importance of one parameter against another (assigning a parameter with a vanishingly small weight effectively eliminates it from the parameter set).

The pressing question, then, is: how do we arrive at a sparse parameter set to define a phonetic space? To achieve temporal sparsity, i.e. to decrease temporal resolution, a natural strategy is to sample sparsely—only at critical points in the timecourse. Which points qualify as critical? For our tritone stimulus set, the humans needed at least 3 samples per syllable to identify degraded stimuli with accuracy approaching that of the intact stimuli, so we could say that the critical points are at syllable onset, midpoint, and offset. But there are alternative criteria. Here, autosegmental-metrical intonational phonology (e.g. Bruce, 1977; Ladd, 2008; Pierrehumbert, 1980) and "landmark" approaches to speech recognition (e.g. Jansen & Niyogi, 2008; Salomon, Espy-Wilson, & Deshmukh, 2004, Stevens, 2002) have converged on the same answer: the critical points are local extrema (peaks and valleys) in the timecourse.[15] Taken together, these different lines of research offer collective insight into temporal aspects of phonetic spaces.

In the autosegmental-metrical theory of intonation, it is the critical or "turning" points in the f0 contour are taken to be the phonetic reflexes of phonological level tonal targets (Ladd, 2008, Section 4.1.2), and transitions between tonal targets are generally specified by linear interpolation (Bruce, 1977; Pierrehumbert & Beckman, 1988) or other mathematical relations, e.g. Pierrehumbert (1981). Evidence that the f0 movements between the tonal targets are not a source of primary parameters defining the phonetic space for intonational tones has come from the phenomenon of "segmental anchoring" (Arvaniti, Ladd, & Mennen, 1998), i.a., see D'Imperio (2012) for a review: that the timing of tonal targets has been shown to consistently be anchored to a linguistic unit, e.g. the stressed syllable onset or nucleus, despite variation in the composition and (absolute) duration of the segmental material between the tonal targets. This work on how tones and segments are co-ordinated draws attention to the choice of temporal unit for the "clock" run to sample parameters from the speech signal. In Section 1.1, we already mentioned that automatic tonal recognition has been striking for its use of coarse sampling resolution, in comparison to automatic speech recognition in general. Another way in which it has been striking is that sampling is often run on a clock defined by linguistic units rather than absolute time, e.g. syllables (Gauthier et al., 2007; Odejobi, Wong, & Beaumont, 2008), the rime (Qian et al., 2007), or the nucleus (Zhou et al., 2008). One way, then, to sample only at critical points would be to sample the f0 contour according to a (language/variety-specific) clock running in temporal units based on the segmental landmarks that tones are anchored to.

Another strategy comes from landmark approaches to speech recognition. Jansen and Niyogi (2008)'s approach involves the integration of information from a collection of independent detectors for different distinctive features, e.g. a stop detector, a nasal detector, etc. A classifier (e.g. an SVM) is built for each distinctive feature, and a detection event of the feature, i.e. when the detector "fires", is defined as a timepoint at which a local maximum in the output of the classifier occurs. The resulting representation of the speech signal is sparse because only timepoints at which a firing event occurs are included—rather than a vector of tens of acoustic parameters every 10 ms, as in many speech recognition systems. The firing events are then integrated at the time scale of the syllable, based on the output of vowel detectors. Jansen and Niyogi (2008)'s landmark approach suggests a more general alterna-

[14] One class of exceptions comes from durational measures, e.g. syllable duration, which has been shown to influence the perception of tonal contrasts (Blicher, Diehl, & Cohen, 1990; Gandour & Harshman, 1978).

[15] Local maxima and minima are in fact "critical points" in the mathematical sense, since they are points at which the first derivative vanishes. In autosegmental-metrical theory, "elbows" in L-shaped f0 curves may also be considered to be critical points; these are points of maximum curvature, i.e. approximately, points where the second derivative is maximized.

tive to sample than running all parameter extraction on a common clock, e.g. sampling each parameter at syllable onset and offset. The idea would be to build independent detectors for different tonal events, e.g. a detector for high tones, another for low tones, as well as detectors for different temporal integration units, e.g. one for syllables, another for utterances. This wouldn't require parameters to be sampled according to some fixed clock, but would allow integration of detector output according to different segmental "anchors" and timescales, as well as classifiers specifically optimized for individual tonal events. For instance, a high tone classifier could be defined over quadratic polynomial coefficients parameterizing the f0 curve, while a low tone classifier might include parameters for vocal fry detection, e.g. Ishi, Ishiguro, and Hagita (2005, 2008).

The importance of integration of detector information over temporal intervals, rather than emphasis on particular instants in time, is also in line with work by Barnes et al. (2012), which argues for the Tonal Center of Gravity model to account for evidence that even while the turning point is held constant, differences in f0 contour shape can still affect the perception of intonational contrasts. In this model, it is not the precise location of turning points in the f0 contour, but the average of timepoints over the time interval of interest, weighted by their measured f0 values, that is used to compute the perceived timing of a tonal event.

## 5. Conclusion

This study adds to the small amount of work that tests the effect of temporal resolution in the speech signal on tone perception. In line with previous work on Mandarin (Gottfried & Suiter, 1997; Lee, 2009; Lee et al., 2008, 2009), it shows that just a few samples per syllable are enough for both humans and support vector machines to classify Cantonese tones with reasonable accuracy, without much difference in performance from having the full speech signal available.

It is important to note two things, though. First, these results are about temporal resolution for *off-line* tonal classification, and not on-line tone processing. Eye movements from online processing of Mandarin tones have shown that tonal perception is an incremental process (Shen, Deutsch, & Rayner, 2013), and as mentioned in Section 1.1, electroencephalographic studies of tone processing also clearly demonstrates that details of contour shape rather than just linear segments interpolating a few points are linguistically encoded in the definition of tonal categories (Chandrasekaran et al., 2007; Krishnan et al., 2009, 2014, 2015, 2017).

Secondly, a closer look at the results from our study shows that even for off-line classification, *where* in the syllable samples are taken greatly impacts how informative they are for tonal identification. This points to further exploration of adaptive sampling, where temporal resolution is not fixed, but may depend on the shape of the f0 contour or other properties of the speech signal; e.g. landmark approaches which extract different parameters from the speech signal on different time scales, determined by when inflection points and local extrema occur in the time course (Jansen & Niyogi, 2009) (see Section 4.3), or approaches where sampling is dependent on the spectral stability of the speech signal (House, 1990, 2004a, 2004b).

The importance of where samples are taken also shows that informative properties of the speech signal to identify a tone may extend to a time window beyond the syllable that the tone is associated to. This has long been established from studies on effects of preceding context on perception, such as Wong and Diehl (2003), which shows that the identification of a word uttered at a given f0 is completely determined by the f0 height of the preceding syllable. But here, we show that information we might consider "intrinsic" to the syllable actually drifts into the following syllable due to peak delay, so the following syllable is part of the domain of tonal realization rather than a source of external contextual information. Thus, in phonetic descriptions of tone, studies of tonal dispersion, and modeling the learning of tones in connected speech, in addition to careful consideration of what temporal resolution to use, we should also consider a time window for parameterization that is larger than just the syllable a tone is associated to.

On a final note, we emphasize again that the role of temporal resolution in phonetic spaces is an issue that transcends which phonetic property or concept is under study. All the issues on temporal resolution of f0 contours we have raised here also apply to any other time-course properties in tonal spaces, e.g. spectral balance and amplitude. And as a more distant case in point, we remind the reader of the example of vowel spaces raised in Section 1: although there is evidence that formant trajectories are important in human vowel perception, results from a large swath of literature in vowel recognition, dispersion, and learnability/acquisition are predicated on the definition of vowel categories using formant values sampled from just a single point in the vowel.

In conclusion, we hope to have shown the reader that valuable insights can come from cross-talk between engineers, neuroscientists, and phoneticians working on similar problems, and to have convinced the reader that temporal resolution in the speech signal is not a finicky technical detail or an engineering problem, but a fundamental issue for phonetics that merits attention from phoneticians.

## Acknowledgements

We wish to acknowledge Hiu Wai Lam for piloting and recording the stimuli and testing the participants in Hong Kong, Eric Zee for so kindly allowing us to test participants in his laboratory at the City University of Hong Kong, Wai Ting Lam for testing participants in Los Angeles, Keelan Evanini for help with implementing RAPT, and Edward Stabler, Megha Sundara and also Patricia Keating, John Kingston, Mark Liberman, and Colin Wilson for illuminating discussions, and three anonymous referees for valuable comments and suggestions. An earlier version of this work was presented at the conference on The Psycholinguistic Representation of Tone in Hong Kong, China in August 2011, and this paper is based in part on the author's Ph.D. dissertation work at the University of California, Los Angeles. This research was supported by an NSF graduate fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## Appendix A. Raw and transformed f0 ranges for speakers of experimental stimuli

The calculated raw and transformed f0 range of the stimuli for each of the five speakers is given in Table A. 1.

**Table A. 1**
Speaker-specific f0 range in speech materials, measured in Hz, after log-transformation, and after standardization of log f0 with respect to speaker means and standard deviations. The speakers are ordered from highest to lowest maximum f0, following the same order from top to bottom in the plot of f0 contours by speaker in Fig. 2 in the paper.

| Speaker | f0 (Hz) | log f0 | z-score |
|---|---|---|---|
| f4 | [165.89, 241.00] | [5.11, 5.48] | [−3.35, 2.35] |
| f3 | [106.42, 179.47] | [4.67, 5.19] | [−5.78, 1.83] |
| m6 | [125.88, 176.36] | [4.84, 5.15] | [−2.92, 3.21] |
| m1 | [83.87, 145.92] | [4.43, 4.97] | [−3.48, 1.84] |
| m5 | [61.44, 140.20] | [4.12, 4.79] | [−5.08, 3.60] |

## Appendix B. Background on support vector machines

We sketch a geometrical characterization of how support vector machines work for the binary case, e.g. for two tone classes, following Bennett and Bredensteiner (2000). Call the two classes Tone A and Tone B. Each stimulus is parameterized as a real-valued $p$-dimensional vector and labeled as belonging to either Tone A or B. Thus, the Tone A and B stimuli sets each comprise a set of points in $\mathbb{R}^p$. The SVM algorithm is a way to determine an optimal decision rule to assign a tone class label to a stimulus. A linear SVM determines a $p-1$ dimensional separating hyperplane as a decision boundary in the parameter space, i.e. a 1-dimensional line for stimuli parameterized in 2-D space, $\mathbb{R}^2$. The SVM algorithm chooses the optimal separating hyperplane to be the one that maximizes the distance from the hyperplane to the Tone A and Tone B sets.

Which hyperplane is this? Take the convex hulls of the Tone A and Tone B sets, the set of points enclosed in the tightest rubber band one can stretch around the Tone A and B sets, respectively. The optimal hyperplane bisects and is orthogonal to the line segment between the two closest points of the convex hulls (Boyd & Vandenberghe, 2004, p. 46–49). If Tone A and B are linearly inseparable, i.e. if their convex hulls overlap, then a soft margin SVM algorithm can be used, which allows for some points to be on the wrong side of the margin in determining the optimal separating hyperplane, and a soft margin parameter is tuned to balance the tradeoff between maximizing the margin and minimizing classification error.

We desire the determined classification rule to generalize beyond the particular set of training data used to choose it. Thus, evaluation of classifier performance, e.g. how accurately it identifies tones, is done by determining classification accuracy on test data, data not in the training data set: in this study, we trained five different classifiers, each one on stimuli from one of the five speakers, and tested the classifiers on the four withheld speakers. For each classifier, we also chose the soft margin parameter to be the value yielding the highest classification accuracy from a grid search over a set of values ranging from $1e^{-2}$ to $1e^{2}$.

## Appendix C. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.wocn.2017.06.004.

## References


Alexander, J. A. (2010). *The theory of adaptive dispersion and acoustic-phonetic properties of cross-language lexical-tone systems* (Ph.D. thesis). Northwestern University.

Andruski, J. E. (2006). Tone clarity in mixed pitch/phonation-type tones. *Journal of Phonetics, 34*, 388–404.

Andruski, J. E., & Costello, J. (2004). Using polynomial equations to model pitch contour shape in lexical tones: An example from Green Mong. *Journal of the International Phonetic Association, 34*, 125–140.

Andruski, J. E., & Ratliff, M. (2000). Phonation types in production of phonological tone: The case of Green Mong. *Journal of the International Phonetic Association, 30*, 37–61.

Arvaniti, A., Ladd, D. R., & Mennen, I. (1998). Stability of tonal alignment: The case of greek prenuclear accents. *Journal of Phonetics, 26*, 3–25.

Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. (2012). Tonal center of gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology, 3*, 337–383.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.

Barry, J. G., & Blamey, P. J. (2004). The acoustic analysis of tone differentiation as a means for assessing tone production in speakers of Cantonese. *The Journal of the Acoustical Society of America, 116*, 1739–1748.

Bashford, J. A., Riener, K. R., & Warren, R. M. (1992). Increasing the intelligibility of speech through multiple phonemic restorations. *Perception and Psychophysics, 51*, 211–217.

Bashford, J. A., Warren, R. M., & Brown, C. A. (1996). Use of speech-modulated noise adds strong bottom-up cues for phonemic restoration. *Perception & Psychophysics, 58*, 342–350.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-6.

Bauer, R. S., Kwan-hin, C., & Pak-man, C. (2003). Variation and merger of the rising tones in Hong Kong Cantonese. *Language Variation and Change, 15*, 211–225.

Becker-Kristal, R. (2010). *Acoustic typology of vowel inventories and dispersion theory: Insights from a large cross-linguistic corpus* (Ph.D. thesis). University of California Los Angeles.

Bennett, K. P., & Bredensteiner, E. J. (2000). Duality and geometry in SVM classifiers. In *Proceedings of the seventeenth international conference on machine learning* (pp. 57–64). Morgan Kaufmann Publishers Inc.

Bishop, J., & Keating, P. (2012). Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex. *Journal of the Acoustical Society of America, 132*, 1100–1112.

Blicher, D. L., Diehl, R. L., & Cohen, L. B. (1990). Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: Evidence of auditory enhancement. *Journal of Phonetics, 18*, 37–49.

Boersma, P., & Weenink, D. (2010). Praat: Doing phonetics by computer (version 5.1.32) [computer program]. http://www.praat.org.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 443–446.

Bruce, G. (1977). *Swedish word accents in sentence perspective*. Lund: CWK Gleerup.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*, 121–167.

Chandrasekaran, B., Krishnan, A., & Gandour, J. T. (2007). Experience-dependent neural plasticity is sensitive to shape of pitch contours. *NeuroReport, 18*, 1963–1967.

Chang, C. -C., & Lin, C. -J. (2001). LIBSVM: A library for support vector machines. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

Chao, Y.-R. (1930). A system of tone-letters. *Le Maître Phonétique, 45*, 24–27.

Chao, Y. R. (1968). *A grammar of spoken Chinese*. Berkeley, CA: University of California Press.

Chen, M., Yang, Z., & Liu, W. (2014). Deep neural networks for Mandarin tone recognition. In *2014 International Joint Conference on Neural Networks (IJCNN)* (pp. 1154–1158).

Ciocca, V., & Bregman, A. S. (1987). Perceived continuity of gliding and steady-state tones through interrupting noise. *Perception and Psychophysics, 42*, 476–484.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America, 116*, 3647–3658.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.

Dannenbring, G. L. (1976). Perceived auditory continuity with alternately rising and falling frequency transitions. *Canadian Journal of Psychology, 30*, 99–114.

de Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online, 4*, 129–134.

de Leeuw, J., & Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software, 31*, 1–30.

De Looze, C., & Rauzy, S. (2009). Automatic detection and prediction of topic changes through automatic detection of register variations and pause duration. In *INTERSPEECH-2009* (pp. 2919–2922).

DiCanio, C., Amith, J. D., & Garća, R. C. (2014). The phonetics of moraic alignment in Yoloxóchitl Mixtec. In *TAL-2014* (pp. 203–210).

Dillon, B., Dunbar, E., & Idsardi, W. (2013). A single-stage approach to learning phonological categories: Insights from Inuktitut. *Cognitive Science, 37*, 344–377.

D'Imperio, M. (2012). Tonal alignment. In A. C. Cohn, C. Fougeron, & M. K. Huffman (Eds.), *The Oxford handbook of laboratory phonology* (pp. 275–287). Oxford, England: Oxford University Press.
Evanini, K. (2011). Improved measurement point selection for automatic formant extraction. In *The 85th annual meeting of the linguistic Society of America. Pittsburgh, PA.* .
Evanini, K., & Lai, C. (2010). The importance of optimal parameter setting for pitch extraction. *The Journal of the Acoustical Society of America, 128*, 2291.
Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review, 120*, 751–778.
Fu, Q.-J., & Zeng, F.-G. (2000). Identification of temporal envelope cues in Chinese tone recognition. *Asia Pacific Journal of Speech, Language and Hearing, 5*, 45–57.
Fu, Q., Zeng, F., Shannon, R. V., & Soli, S. D. (1998). Importance of tonal envelope cues in Chinese speech recognition. *The Journal of the Acoustical Society of America, 104*, 505–510.
Gandour, J. (1981). Perceptual dimensions of tone: Evidence from Cantonese. *Journal of Chinese Linguistics, 9*, 20–36.
Gandour, J. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics, 11*, 149–175.
Gandour, J., Potisuk, S., Dechonkit, S., & Ponglorpisit, S. (1992). Tonal coarticulation in Thai disyllabic utterances: A preliminary study. Linguistics of the Tibeto-Burman area, 15.
Gandour, J. T., & Harshman, R. A. (1978). Crosslanguage differences in tone perception: A multidimensional scaling investigation. *Language and Speech, 21*, 1–33.
Ganong, W. F. III, (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*, 110–125.
Garellek, M., Keating, P., Esposito, C. M., & Kreiman, J. (2013). Voice quality and tone identification in White Hmong. *The Journal of the Acoustical Society of America, 133*, 1078–1089.
Gauthier, B., Shi, R., & Xu, Y. (2007). Learning phonetic categories by tracking movements. *Cognition, 103*, 80–106.
Gottfried, T. L., & Suiter, T. L. (1997). Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones. *Journal of Phonetics, 25*, 207–231.
Grassi, M., & Soranzo, A. (2009). MLP: A MATLAB toolbox for rapid and reliable auditory threshold estimation. *Behavior Research Methods, 41*, 20–28.
Greenberg, S., & Zee, E. (1977). On the perception of contour tones. *UCLA Working Papers in Phonetics, 45*, 150–159.
Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.
Han, M., & Kim, K.-O. (1974). Phonetic variation of Vietnamese tones in disyllabic utterances. *Journal of Phonetics, 22*, 477–492.
Hant, J. J., & Alwan, A. (2003). A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise. *Speech Communication, 40*, 291–313.
Hermes, D. J. (2006). Stylization of pitch contours. In S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, & I. Mleinek, et al. (Eds.), *Methods in empirical prosody research* (pp. 29–62). Walter de Gruyter.
Hess, W. (1983). *Pitch determination of speech signals: Algorithms and devices*. Springer-Verlag.
Hirst, D., & Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix, 15*, 71–85.
House, D. (1990). *Tonal perception in speech*. Lund, Sweden: Lund University Press.
House, D. (2004a). Pitch and alignment in the perception of tone and intonation. In G. Fant, H. Fujisaki, J. Cao, & Y. Xu (Eds.), *From traditional phonology to modern speech processing* (pp. 189–204). Foreign Language Teaching and Research Press.
House, D. (2004b). Pitch and alignment in the perception of tone and intonation: Pragmatic signals and biological codes. In *International symposium on tonal aspects of languages with emphasis on tone languages* (pp. 93–96).
Ishi, C. T., Ishiguro, H., & Hagita, N. (2005). Proposal of acoustic measures for automatic detection of vocal fry. In *INTERSPEECH-2005* (pp. 481–484).
Ishi, C. T., Sakakibara, K.-I., Ishiguro, H., & Hagita, N. (2008). A method for automatic detection of vocal fry. *IEEE Transactions on Audio, Speech, and Language Processing, 16*, 47–56.
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York, NY: Springer.
Jansen, A., & Niyogi, P. (2008). Modeling the temporal dynamics of distinctive feature landmark detectors for speech recognition. *The Journal of the Acoustical Society of America, 124*, 1739–1758.
Jansen, A., & Niyogi, P. (2009). Point process models for event-based speech recognition. *Speech Communication, 5*, 1155–1168.
Khouw, E., & Ciocca, V. (2007). Perceptual correlates of Cantonese tones. *Journal of Phonetics, 35*, 104–117.
Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America, 118*, 1038–1054.
Krishnan, A., Gandour, J. T., & Ananthakrishnana, S. (2014). Cortical pitch response components index stimulus onset/offset and dynamic features of pitch contours. *Neuropsychologia, 59*, 1–12.
Krishnan, A., Gandour, J. T., Ananthakrishnana, S., & Vijayaraghavan, V. (2015). Language experience enhances early cortical pitch-dependent responses. *Journal of Neurolinguistics, 33*, 128–148.
Krishnan, A., Gandour, J. T., Bidelman, G. M., & Swaminathan, J. (2009). Experience-dependent neural representation of dynamic pitch in the brainstem. *NeuroReport, 20*, 408–413.
Krishnan, A., Gandour, J. T., Xu, Y., & Suresh, C. H. (2017). Language-dependent changes in pitch-relevant neural activity in the auditory cortex reflect differential weighting of temporal attributes of pitch contours. *Journal of Neurolinguistics, 41*, 38–49.
Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General, 142*, 573–603.
Kuang, J. (2013). The tonal space of contrastive five level tones. *Phonetica, 70*, 1–23.
Labov, W., Ash, S., & Boberg, C. (2006). *The Atlas of North American English*. Berlin, Germany: Mouton de Gruyter.
Ladd, D. R. (2008). *Intonational phonology* (2nd ed.). Cambridge University Press.
Ladefoged, P. (1980). What are linguistic sounds made of? *Language, 56*, 485–502.
Lam, Y. F. (2014). *Cantonese tone recognition using the Hilbert-Huang transform* Master's thesis. China: The Hong Kong University of Science and Technology Hong Kong.
Laniran, Y. O. (1992). *Intonation in tone languages: The phonetic implementation of tones in Yoruba* (Ph.D. thesis). Cornell University.
Lee, C. (2009). Identifying isolated, multispeaker Mandarin tones from brief acoustic input: A perceptual and acoustic study. *The Journal of the Acoustical Society of America, 125*, 1125–1137.
Lee, C., Tao, L., & Bond, Z. (2008). Identification of acoustically modified Mandarin tones by native listeners. *Journal of Phonetics, 36*, 537–563.
Lee, C., Tao, L., & Bond, Z. (2009). Speaker variability and context in the identification of fragmented Mandarin tones by native and non-native listeners. *Journal of Phonetics, 37*, 1–15.
Levow, G.-A. (2005). Context in multi-lingual tone and pitch accent recognition. In *Proceedings of INTERSPEECH 2005* (pp. 1809–1812).
Levow, G.-A. (2006). Unsupervised and semi-supervised learning of tone and pitch accent. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL* (pp. 224–231).
Li, Q., & Chen, Y. (2016). An acoustic study of contextual tonal variation in Tianjin Mandarin. *Journal of Phonetics, 54*, 123–150.
Li, Y., & Lee, T. (2007). Perceptual equivalence of approximated Cantonese tone contours. In *INTERSPEECH-2007* (pp. 2677–2680).
Li, Y., & Lee, T. (2008). A perceptual study of approximated Cantonese tone contours. In *2008. ISCSLP '08. 6th international symposium on Chinese spoken language processing* (pp. 1–4).
Li, Y., Lee, T., & Qian, Y. (2002). Acoustical F0 analysis of continuous Cantonese speech. In *Proceedings of international symposium on Chinese spoken language processing* (pp. 127–130).
Li, Y., Lee, T., & Qian, Y. (2004). F0 analysis and modeling for Cantonese Text-to-Speech. In *SP-2004* (pp. 467–470).
Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language, 48*, 839–862.
Lisker, L. (1978). Rapid vs rabid: A catalogue of acoustic features that may cue the distinction. *Haskins Laboratories Status Report on Speech Research, SR-54*, 127–132.
Liu, S., & Samuel, A. G. (2004). Perception of Mandarin lexical tones when f0 information is neutralized. *Language and Speech, 47*, 109–138.
Luce, R. D. (1963). Detection and recognition. In R. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 103–189). Wiley.
Matthews, S., & Yip, V. (1994). *Cantonese: A comprehensive grammar*. New York: Routledge.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2012). e1071: Misc functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-1.
Miller, G. A., & Licklider, J. C. R. (1950). The intelligibility of interrupted speech. *The Journal of the Acoustical Society of America, 22*, 167–173.
Mok, P. P.-K., & Wong, P. W.-Y. (2010a). Perception of the merging tones in Hong Kong Cantonese: Preliminary data on monosyllables. In *INTERSPEECH-2010*. .
Mok, P. P.-K., & Wong, P. W.-Y. (2010b). Production of the merging tones in Hong Kong Cantonese: Preliminary data on monosyllables. In *INTERSPEECH-2010*. .
Morén, B., & Zsiga, E. (2006). The lexical and post-lexical phonology of Thai tones. *Natural Language & Linguistic Theory, 24*, 113–178.
Nearey, T. M., & Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification. *The Journal of the Acoustical Society of America, 80*, 1297–1308.
Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology, 34*, 393–418.
Odejobi, O. A., Wong, S. H. S., & Beaumont, A. J. (2008). A modular holistic approach to prosody modelling for Standard Yorùbá speech synthesis. *Computer Speech & Language, 22*, 39–68.
Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437–442.
Peng, G., & Wang, W. S. (2005). Tone recognition of continuous Cantonese speech based on support vector machines. *Speech Communication, 45*, 49–62.
Peng, G., Zheng, H., & Wang, W. S. Y. (2004). Tone recognition for Chinese speech: A comparative study of Mandarin and Cantonese. In *Chinese spoken language processing, 2004 international symposium on* (pp. 233–236).
Peperkamp, S., Calvez, R. L., Nadal, J., & Dupoux, E. (2006). The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition, 101*, B31–B41.
Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America, 24*, 175–184.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation* (Ph.D. thesis). MIT.
Pierrehumbert, J. (1981). Synthesizing intonation. *Journal of the Acoustical Society of America, 70*, 985–995.
Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech, 46*, 115–154.
Pierrehumbert, J., & Beckman, M. (1988). *Japanese tone structure*. The MIT Press.
Pisarn, C., & Theeramunkong, T. (2006). Improving Thai spelling recognition with tone features. In T. Salakoski, F. Ginter, S. Pyysalo, & T. Pahikkala (Eds.), *Advances in natural language processing: 5th international conference on NLP, FinTAL 2006 Turku, Finland, August 23–25, 2006 proceedings* (pp. 388–398). Berlin, Heidelberg: Springer, Berlin Heidelberg.
Prom-on, S., Xu, Y., & Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America, 125*, 405–424.
Prukkanon, N., Chamnongthai, K., & Miyanaga, Y. (2016). F0 contour approximation model for a one-stream tonal word recognition system. *AEU – International Journal of Electronics and Communications, 70*, 681–688.
Qian, Y., Lee, T., & Soong, F. K. (2007). Tone recognition in continuous Cantonese speech using supratone models. *The Journal of the Acoustical Society of America, 121*, 2936–2945.
R Core Team (2014). *R: A language and environment for statistical computing*. Austria: R Foundation for Statistical Computing Vienna.
Reddy, S., & Stanford, J. (2015). Toward completely automated vowel extraction: Introducing DARLA. *Linguistics Vanguard, 1*, 15–28.
Remijsen, B. (2013). Tonal alignment is contrastive in falling contours in Dinka. *Language, 89*, 297–327.
Remijsen, B., & Ayoker, O. G. (2014). Contrastive tonal alignment in falling contours in Shilluk. *Phonology, 31*, 435–462.
Rosenfelder, I., Fruehwald, J., Evanini, K., & Yuan, J. (2011). FAVE (forced alignment and vowel extraction) program suite. http://fave.ling.upenn.edu.
Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., & Prichard, H. (2014). FAVE (Forced Alignment and Vowel Extraction) Program Suite v1.2.2 10.5281/zenodo.22281.
Salomon, A., Espy-Wilson, C. Y., & Deshmukh, O. (2004). Detection of speech landmarks: Use of temporal information. *Journal of the Acoustical Society of America, 115*, 1296–1305.
Samuel, A. (1996). Phoneme restoration. *Language and Cognitive Processes, 11*, 647.
Shen, J., Deutsch, D., & Rayner, K. (2013). On-line perception of Mandarin Tones 2 and 3: Evidence from eye movements. *Journal of the Acoustical Society of America, 133*, 3016–3029.
Shih, C., & Lu, H.-Y. D. (2015). Effects of talker-to-listener distance on tone. *Journal of Phonetics, 51*, 6–35. What's So Special About H(igh)? Multi-Disciplinary Perspectives on the Linguistic Functions of Raised Pitch.
Silbert, N. (2014). Visualizing confusion matrices. http://www.nhsilbert.net/source/2014/03/visualizing-confusion-matrices/.
Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America, 111*, 1872–1891.
Strange, W., Jenkins, J. J., & Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. *The Journal of the Acoustical Society of America, 74*, 695–705.
Sumner, M., & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language, 60*, 487–501.
Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In W. B. Kleijn & K. K. Paliwal (Eds.), *Speech coding and synthesis* (pp. 495–518). Elsevier Science Inc.
Taylor, P. (2000). Analysis and synthesis of intonation using the Tilt model. *The Journal of the Acoustical Society of America, 107*, 1697–1714.
Tian, Y., Zhou, J., Chu, M., & Chang, E. (2004). Tone recognition with fractionized models and outlined features. In *Acoustics, speech, and signal processing, 2004. Proceedings. (ICASSP '04). IEEE international conference on* (Vol. 1). , pp. I–105–8.
UCLA Statistical Consulting Group (2011). R library: Contrast coding systems for categorical variables. Accessed July 22, 2014.
Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences, 104*, 13273–13278.
Vapnik, V. N. (1995). *The nature of statistical learning*. Springer.
Wang, S., & Levow, G.-A. (2008). Mandarin Chinese tone nucleus detection with landmarks. In *Proceedings of interspeech* (pp. 1101–1104).
Wang, S., Tang, Z., Zhao, Y., & Ji, S. (2009). Tone recognition of continuous Mandarin speech based on binary-class SVMs. In *Information science and engineering (ICISE), 2009 1st international conference on* (pp. 710–713).
Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167*, 392–393.
Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica, 49*, 25–47.
Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Springer.
Wong, Y. W. (2006a). Contextual tonal variations and pitch targets in Cantonese. In *Proceedings of speech prosody 2006, Dresden.* .
Wong, Y. W. (2006b). Realization of Cantonese rising tones under different speaking rates. In *Proceedings of speech prosody 2006, Dresden.* .
Wong, P. C. M., & Diehl, R. L. (2003). Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *Journal of Speech, Language & Hearing Research, 46*, 413–421.
Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics, 25*, 61–83.
Xu, Y. (1998). Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica, 55*, 179–203.
Xu, Y. (2001). Fundamental frequency peak delay in Mandarin. *Phonetica, 58*, 26–52.
Yu, K. M., & Lam, H. W. (2011). The role of creaky voice in Cantonese tonal perception. In *Proceedings of ICPhS XVII.* .
Yu, K. M., & Lam, H. W. (2014). The role of creaky voice in Cantonese tonal perception. *Journal of the Acoustical Society of America, 136*, 1320–1333.
Zhang, J., & Hirose, K. (2000). Anchoring hypothesis and its application to tone recognition of chinese continuous speech. In *Acoustics, speech, and signal processing, 2000. ICASSP '00. Proceedings. 2000 IEEE international conference on* (Vol. 3, pp. 1419–1422). .
Zhang, J., & Hirose, K. (2004). Tone nucleus modeling for Chinese lexical tone recognition. *Speech Communication, 42*, 447–466.
Zhou, N., Zhang, W., Lee, C., & Xu, L. (2008). Lexical tone recognition with an artificial neural network. *Ear and Hearing, 29*, 326–335. PMC2562432.
Zsiga, E., & Nitisaroj, R. (2007). Tone features, tone perception, and peak alignment in thai. *Language and Speech, 50*, 343–383.