10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Intermediate features are not useful for tone perception

Yue Chen, Yi Xu

Department of Speech, Hearing and Phonetic Sciences, University College London, London, United Kingdom

yue.chen.1@ucl.ac.uk, yi.xu@ucl.ac.uk

Abstract

Many theories assume that speech perception is done by first extracting features like the distinctive features, tonal features or articulatory gestures before recognizing phonetic units such as segments and tones. But it is unclear how exactly extracted features can lead to effective phonetic recognition. In this study we explore this issue by using support vector machine (SVM), a supervised machine learning model, to simulate the recognition of Mandarin tones from F_0 in continuous speech. We tested how well a five-level system or a binary distinctive features system can identify Mandarin tones by training the SVM model with F_0 trajectories with reduced temporal and frequency resolutions. At full resolution, the recognition rates were 97% and 86% based on the semitone and Hertz scales, respectively. At reduced temporal resolution, there was no clear decline in recognition rate until two points per syllable. At reduced frequency resolution, the recognition rate dropped rapidly: by the level with 5 bands, the accuracy was around 40% based on both Hertz and semitone scales. These results suggest that intermediate featural representations provide no benefit for tone recognition, and are unlikely to be critical for tone perception.

Index Terms: speech perception, distinctive features, tone, SVM

1. Introduction

How exactly humans perceive speech is still a mystery. It is widely assumed that multiple acoustic cues are needed for the perception of segmental and suprasegmental units, and a major goal of phonetics is to find out which cues are relevant for the recognition of phonetic categories [1], [2]. For example, formants may provide primary cues for vowel categories; VOT is useful for distinction between voiced and voiceless plosives; pitch contours are useful for differentiating lexical tones, etc. Those cues are then combined to identify phonetic or phonemic categories. What has rarely been contemplated, however, is exactly how those cues are used to achieve the recognition of those categories from continuous speech. In this paper we explore the idea that speech perception cannot be adequately understood unless the role of features and cues is reconsidered.

A major source of the cue-based view of speech perception is the classic theory of distinctive features [3]. The theory started as an attempt to economize the representation of phoneme categories by minimizing the number of acoustic properties needed to classify segments [3]. A system was proposed with only 12 features = pairs, each for making a binary contrast based predominantly on acoustic properties [4], [5]. Subsequently, an alternative system was proposed by Chomsky and Halle with a much larger set of binary features which are predominantly based on articulatory properties [6]. Further development of the feature theory also attempted to relax the binary assumption by allowing multivalued features [6], [7]. Consistent with the feature theory is the widely accepted idea that feature or cue extraction is the key to speech perception [8], [9]. This is true of both the auditory theory and motor theory of speech perception, two competing frameworks that have been dominating this area of research.

The auditory theory assumes that it is the auditory properties of phonetic events that listeners attend to [10]. The motor theory, in contrast, assumes that speech perception is achieved with an articulatory recognition phase before the identification of words [11], [12] and [13]. Both theories, therefore, assume that an intermediate phase is needed in perception in which features are extracted from continuous speech signal. They differ from each other only in terms of whether the extracted features are primarily auditory or articulatory in nature. Both theories, as well as many other conceptual frameworks of speech perception, share in common what we would refer to as the *feature-to-percept* assumption. None of these theories, however, offers proposals about how exactly the features are extracted, or how the detected features are processed to identify or discriminate the sound categories.

There have been attempts to use features in automatic speech recognition systems. For example, the landmark-based approach tries to extract distinctive features from around acoustic landmarks such as consonant closure, which can then be used to make choices between candidate segments [14], [15] and [16]. In most cases, however, systems using landmarks or distinctive features are knowledge-based, and the detected features are used only as one kind of features among other linguistic and acoustic features to recognize phonemes [17], [18]. To our knowledge, there is no speech recognition system fully based on extracted distinctive features. There have also been systems that perform speech recognition with the help of articulatory features, e.g., [18]. But again, we are not aware of any system that can recognize speech segments based on articulatory gestures extracted from acoustic signals.

The distinctive feature theory has also influenced research on lexical tones. Binary tone features were introduced by [19], although its usefulness in phonological analysis has been questioned [20], [21]. A more broadly accepted practice is to use a five-level system proposed by Chao [22]. This system moves away from the binary assumption of the classical feature theory, but it nevertheless assumes that only 5 discrete levels are needed to distinguish all the tones of a language. Also different from the classical feature theory, the five-level system allows representation of pitch change over time by representing each tone with two temporal points. As an example, the four lexical tones of Mandarin are represented as 55—tone1, 35—tone 2, 214—tone 3, and 51—tone 4, where a larger number indicates a higher pitch. These numerical forms are abstracted from the original pitch forms. There are also alternative

10.21437/SpeechProsody.2020-105

schemes that try to represent tone contours [23], [24] and [25], but they also focus on abstracting pitch contours into several discrete levels. What is common to all these tone representation schemes is the implication that to perceive a tone, listeners need to *first* derive at a discrete form of representation, and then use this kind of representation to identify the tone. What has not been considered is the possibility that a tone could be identified by directly processing fully continuous F_0 contours, without extracting any *discrete* representations. In this way, there is no need to determine, e.g., for each of the discrete temporal locations, what is the pitch level that is also discrete.

The aim of this study is to test these two tone perception hypotheses by comparing the performance of computation modeling of tone recognition using either fully continuous F_0 contours, or contours with varying degrees of reduction of resolution. The resolution reduction is a way to simulate feature extraction, based on the assumption that a limited set of discrete levels and temporal locations are sufficient to fully represent the tones of a language. While resolution reduction may not be the best way to simulate feature extraction, there is lack of proposal from those who believe features are necessary on how exactly they can be extracted. The computational model used in this study is SVM.

2. Method

2.1. Support Vector Machine (SVM)

SVM is a supervised machine learning model developed for binary classification tasks involving calculating Euclidean distances in the parameter space. In the application of SVM, all the samples (sequences of n f0) are converted to n-dimensional vectors and labeled. The weight W is a combination of a subset of the training examples and show how each dimension of the vectors is used in the classification process. The vectors (x) in this subset are named support vectors. A simple functional margin can be: $f(x) = sign(W^T x + b)$, where b is a constant. In our experiment, one F_0 contour is one sample consisting of 30 sample points, and treated as a 30-dimention vector. The classification is done with the LibSVM tool [26] with RBF kernel. It generalizes the binary classification to a n-class classifier that splits the task into n(n−1)/2 binary tasks and the solutions are combined by a voting strategy [27].

2.2. Simulation of feature extraction

For any highly abstract features to work, one of the first critical steps is to extract them from observation by identification and naming. This is by no means a trivial task, and its effectiveness can be shown only in terms of the ultimate rate of recognition of the phonetic category. Alternatively, we can partially simulate the effect of feature extraction by reducing signal resolution toward the level specified by particular feature system. For the commonly used 5-level tone representation system mentioned earlier, this would mean to a) reduce temporal resolution to 2 points/tone, and b) reduce frequency resolution to 5 levels/tone.

The data were syllable-sized F_0 contours produced by four female and four male Mandarin speakers [28]. Each stimulus corresponds to a syllable in a disyllabic "mama" produced in a carrier sentences with either high or low pre-f0 value and post-f0 value. Each token is a 30 equidistant (hence time-normalized) discrete points vector taken from either the first or second syllable of a disyllabic tone sequence in the middle position of a carrier sentence. There was no separation of the tokens from the first and second syllables, thus leaving the information of syllable position in word/phrase unrepresented. Two types the raw data were used, F_0 in Hertz and semitones. The latter was converted from Hertz with the following equation:

$$semitone = log_2(f0) * 12 \quad (1)$$

where the reference F_0 is assumed to be 1 Hz for all speakers. Note that this kind of raw data retains most of the individual differences in pitch height, particularly between the female and male speakers, as can be clearly seen in Figure 1 and 2 which show plots of F_0 contours in Hertz and semitone, respectively.

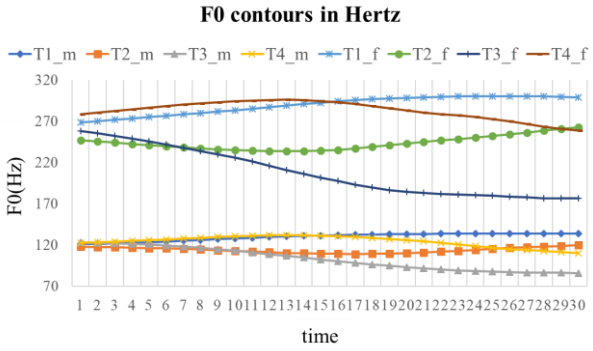

Figure 1: *Mean time-normalized syllable-sized F_0 contours of four Mandairn tones, averaged separately for female and male speakers.*

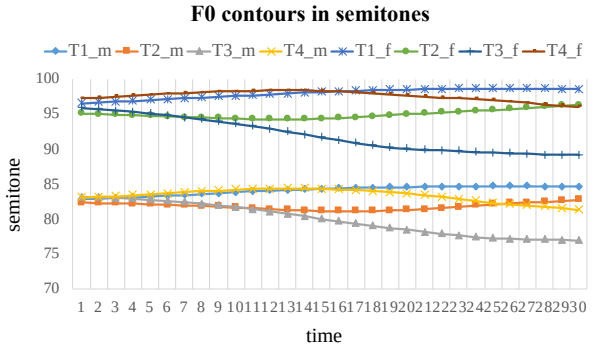

Figure 2: *Mean time-normalized syllable-sized semitone contours of four Mandairn tones, averaged separately for female and male speakers.*

There were a total of 1408 tokens of Tone 1 (high-level), 1408 tokens of Tone2 (rising), 1232 tokens of Tone 3 (low), and 1408 tokens of Tone 4 (falling). The number of tokens was fewer for Tone 3 because we removed those of the first syllable followed by another Tone 3 to circumvent the problem of the well-known tone sandhi rule, which makes the first Tone 3 to closely resemble Tone 2 [22], [28]. The whole dataset was then divided into a training subset and a testing subset, with a ratio of 2:1.

2.2.1. Reduction of temporal resolution

As mentioned before, the widely used featural system of the four Mandarin tones in connected speech (which differs slightly from their canonical forms in isolation) represents each of them with two discrete scales, with 2 temporal specifications and 5 frequency specifications: 55 for Tone1, 35 for Tone 2, 21 for Tone 3, and 51 for Tone 4. Following this scheme, the temporal resolution can be reduced down to 2 points/tone. To give the

feature theory the benefit of the doubt, we have assumed that after extracting the theory-specified number of points, interpolations are applied to link up those points to form intermediate featural contours, as illustrated in Figure 3. These featural contours were then used as the testing data against the tone classifiers trained with full F_0 contours (i.e., those with 30 points/tone on a continuous frequency scale). In order to see how the performance of tone recognition may change gradually as a function of temporal resolution reduction, we also extracted varying number of equidistant points along the F_0 contours. For all reduced temporal resolutions, linear interpolations were applied to link up adjacent points. The newly synthesized samples are the testing data.

2.2.2. *Reduction of frequency resolution*

The reduction of frequency was done by fitting the F_0 points (with or without reduced temporal resolution) into different number of fixed bands to form the testing data. The training data were again the original vectors. And like for the temporal dimension, in order to see how the performance of tone recognition may change gradually as a function of frequency resolution reduction, a varying number of bands were used.

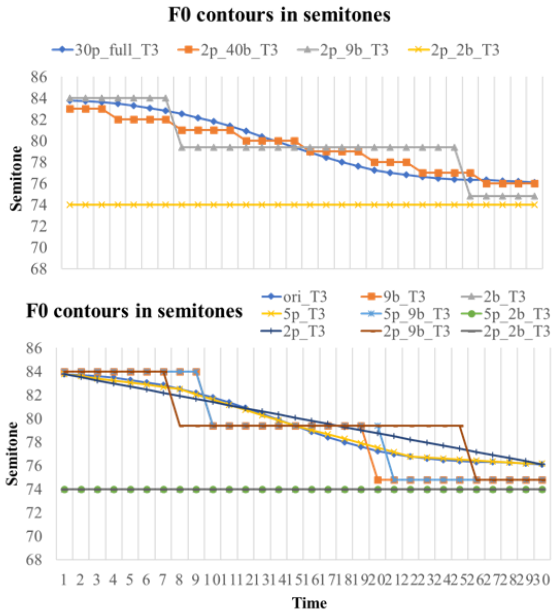

Figure 3: *F_0 contours in semitones with various levels of reduction of temporal and frequency resolutions. (In the legend, np refers to n points and mb refers to m bands after reduction).*

2.3. Experimental set up

The experiment used F_0 contours of the four Mandarin tones with full resolution to train an SVM model, which was then used to classify the tonal categories of contours with varying levels of resolution reduction, as shown in Table 1. Thus, there were three experimental conditions, all with full resolution as the reference condition:

1. Reduced temporal resolution alone;
2. Reduced frequency resolution alone;
3. Reduce temporal and frequency resolution combined.

Table 1: *Levels of resolution reduction.*

No. of temporal points	No. of frequency bands
30	Full
20	≈ 40
15	18
10	9
5	5
3	2
2	

In total, there were 7*6=42 trials for contours in Hertz and 42 trials for contours in semitones.

3. Result

3.1. Effect of reduced temporal resolution

As shown in Figure 4, at full resolution, the error rates were low, at 2.6% for contours in semitones and 16% for contours in Hertz. The lower rate of contours in Hz is unsurprising, as the logarithmic conversion in calculating semitones has effectively normalized the vertical span of the pitch range, with only individual differences in pitch height still retained.

As temporal resolution reduced, the rate of tone recognition remains high until there were only two temporal points left. Note that, two temporal points per tone is exactly what most featural tonal representations assume to be adequate.

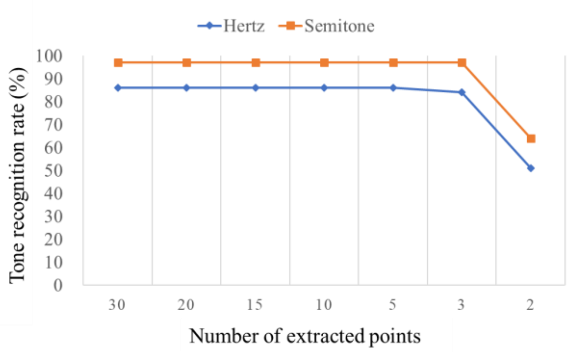

Figure 4: *Tone recognition rate as a function of temporal resolution.*

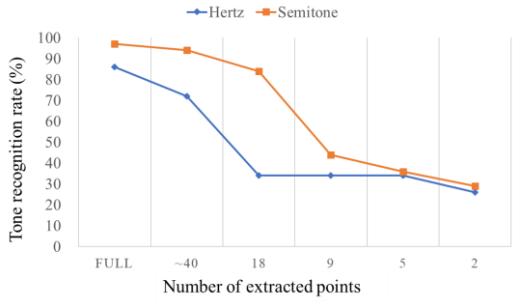

Figure 5: *Tone recognition rate as a function of frequency resolution.*

3.2. Effect of reduced frequency resolution

As shown in Figure 5, tone recognition rate is more sensitive to reduction in frequency resolution than to temporal resolution. It starts to drop at 40 bands, and the decline is faster for contours in Hertz than for those in semitones. By 5 bands, the recognition

rates have dropped below 40% for contours on both frequency scales.

3.3. Effect of reduced temporal and frequency resolution combined

As shown in Figures 6 and 7, when temporal and frequency resolutions were reduced at the same time, only the effect of frequency resolution can be clearly seen, as the recognition rates are largely the same across different temporal resolutions, except for the two-point condition which started low even with full frequency resolution. The only difference between the two figures are that a) a major drop occurred between 37 and 18 bands for contours in Hertz (Figure 6) while a major drop occurred between 18 and 9 bands for contours in semitones (Figure 7), and b) the rate is overall lower for the Hertz scale than the semitone scale.

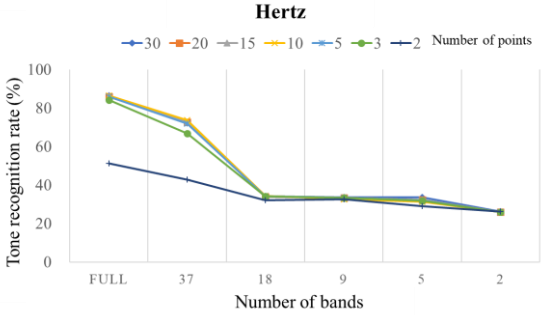

Figure 6: *Tone recognition rate as a function of both temporal and frequency resolution on a Hertz scale.*

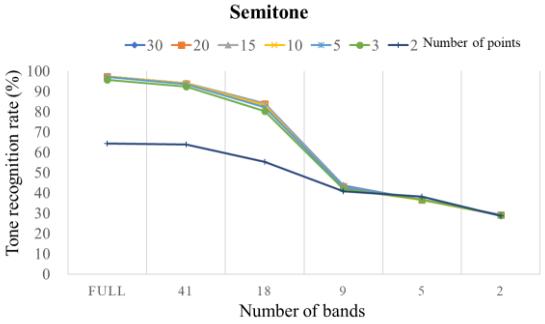

Figure 7: *Tone recognition rate as a function of both temporal and frequency resolution based on a semitone scale*

4. Discussion

The experimental results have shown that at full data resolution, F_0 contours both in Hertz and in semitones can achieve high tone recognition rates (86% and 97%, respectively). Although similar recognition rates were already shown in a previous study using the same corpus [29], the performance is not trivial, as these tones were produced in fluent connected speech in many different tonal contexts and two syllable positions [28], yet in the present study no contextual or positional information is used as input features during training and testing, contrary to the common practice in speech technology applications [30], [31] and [32]. This means that, despite the variability, tones produced in contexts by speakers of both genders still have enough in common to allow a pattern recognition algorithm (SVM) to accurately recognize the tonal categories. More importantly, given that the ultimate goal of tone recognition is the recognition of tones, no intermediate steps are needed to first extract "economical" tonal features before recognizing the tones.

Yet potential benefits of tonal features cannot be ruled out until it is tested, which is preliminarily done in the present study by reducing either temporal resolution or frequency resolution or both from the raw F_0 data. No benefits, however, were seen in terms of improved tone recognition rate. Instead, with reduced temporal resolution, the performance did not change much until the number of F_0 points were reduced to two, when the recognition rates were dropped to 51.2% and 64.2% for F_0 contours in Hertz and semitones, respectively. Interestingly, two pitch specifications per tone is exactly what is assumed by the widely accepted featural representations of Mandarin tones [19], [22] and [33]. With reduced frequency resolution, tone recognition rate dropped quickly. The decline was already apparent from full resolution to around 40 frequency bands for both Hertz and semitone scales. Then a significant drop to below 50% occurred by 18 bands for the Hertz scale and 9 bands for the semitone scale, respectively. For both scales, therefore, the recognition rates are already at a level that no theory would consider as viable. When there are only five frequency bands, i.e., at the level of the widely accepted five-level feature system, the recognition rates were 33% for Hertz and 36% for semitone scales.

It could be argued that the five-level tone scale is supposed to represent the pitch range of individual speakers, and that is why there is often claimed to be a need to first normalize the pitch range of all speakers by applying a Z-score conversion to [34], [35]. But for real life speech perception, one may wonder how listeners can perform such pitch range normalization when they have heard just one or a few utterances from a speaker? Furthermore, if, as shown in the present study and in [29] that tone recognition can be achieved at a high accuracy without speaker normalization as long as full resolution of the raw data is retained, why should there still be a need to extract the features? In fact, even the high tone recognition rate achieved with full resolution in the present study might be better than real life perception, as the stimuli used did not contain adverse factors like consonantal perturbation of F_0 [36], intonational confounds [37], etc. This would make it even more important that as little information as possible is discarded from the raw data in perception before a tone is actually recognized.

5. Conclusion

We have shown that syllable-sized continuous F_0 contours with full resolution can be used to train a SVM model to achieve high tone recognition rates, without extracting intermediate features. We have further demonstrated that reducing temporal and frequency resolution from the contours contour would result in reduced rate of tone recognition. By the time the resolution is equivalent to the widely used five-level, two-point tone representation, the recognition rate had dropped to around 40%. These findings pose serious questions about what we would call the feature-to-percept assumption. As an alternative, we would like to suggest that raw acoustic signals can be processed directly to recognize phonetic categories without explicitly extracting intermediate features. Observations of the role of certain features are useful only for descriptive purposes.

6. References

[1] Ladefoged, Peter, "What are linguistic sounds made of?" *Language*, vol. 56, pp. 485-502, Sept. 1980.

[2] Wright, Richard, "A review of perceptual cues and cue robustness," in *Phonetically based phonology*, B. Hayes et.al, Eds. Cambridge: Cambridge University Press, 2004, pp. 34-57.
[3] Jakobson, Roman, C. Gunnar Fant, and Morris Halle, *Preliminaries to speech analysis: The distinctive features and their correlates*, Cambridge: Massachusetts Institute of Technology, 1951.
[4] Jakobson, Roman, and Moris Halle, *Fundamentals of language*. 2nd ed, Vol. 1, Germany: Walter de Gruyter, 2010.
[5] Jakobson, Roman, and Morris Halle, *Phonology in Relation to Phonteics*. North-Holland Publishing Company, 1968.
[6] Chomsky, Noam, and Morris Halle, *The sound pattern of English*, New York: Harper & Row, 1968.
[7] Ladefoged, Peter, *Preliminaries to linguistic phonetics*, Chicago: University of Chicago Press, 1971.
[8] Flemming, Edward S, *Auditory representations in phonology*. Routledge, 2013.
[9] Kingston, John, and Randy L. Diehl, "Intermediate properties in the perception of distinctive feature values," in *laboratory phonology 4*, B. Connell and A. Arvaniti, Eds. Cambridge: Cambridge University Press, 1995, pp. 7-27.
[10] Diehl, Randy L., and Keith R. Kluender, "On the objects of speech perception," *Ecological psychology*, vol. 1.2, pp. 121-144, 1989.
[11] Liberman, Alvin M., et al, "Perception of the speech code," *Psychological review*, vol. 74.6, pp. 431-461, 1967.
[12] Liberman, Alvin M., and Ignatius G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21.1, pp. 1-36, 1985
[13] Galantucci, Bruno, Carol A. Fowler, and Michael T. Turvey, "The motor theory of speech perception reviewed." *Psychonomic bulletin & review*, vol.13.3, pp. 361-377, 2006.
[14] Stevens, Kenneth N., et al, "Implementation of a model for lexical access based on features," in *Second International Conference on Spoken Language Processing*, Banff, Alberta, Canada, 1992, pp. 499-502.
[15] Deng, Li, and Don Sun, "Speech recognition using the atomic speech units constructed from overlapping articulatory features," in *Third European Conference on Speech Communication and Technology*, Berlin, Germany, 1993, pp. 1635-1638.
[16] Liu, Sharlene A, "Landmark detection for distinctive feature‐based speech recognition," *The Journal of the Acoustical Society of America*, vol. 100.5, pp. 3417-3430, 1996.
[17] Eide, Ellen, "Distinctive features for use in an automatic speech recognition system," in *Seventh European Conference on Speech Communication and Technology*, Scandinavia, 2001.
[18] Erler, Kevin, and George H. Freeman, "An HMM‐based speech recognizer using overlapping articulatory features," *The Journal of the Acoustical Society of America*, vol. 100.4, pp. 2500-2513, 1996.
[19] SY, William, "Phonological features of tone," *International Journal of American Linguistics*, vol. 33.2, pp. 93-105, 1967.
[20] Clements, George N., Alexis Michaud, and Cédric Patin, "Do we need tone features," in *Tones and Features: Phonetic and Phonological Perspectives*, J. Goldsmith, E. Hume and WL. Wetzels, Berlin: De Gruyter Mouton, 2011, pp. 3-24.
[21] Hyman, Larry M, "Do tones have features," in *Tones and Features: Phonetic and Phonological Perspectives*, J. Goldsmith, E. Hume and WL. Wetzels, Berlin: De Gruyter Mouton, 2011, pp. 50-80.
[22] Chao, Yuen Ren, *Language and symbolic systems*, Cambridge: Cambridge University Press, 1968.
[23] Shi F and Liao R, *Essays on Phonetics*, Beijing: Beijing Language and Culture Press, 1994.
[24] Zhu X, *Phonetics*, Shanghai: Commercial Press, 2010.
[25] Zhu X, *Records of Shanghai Tonal Experiments*, Shanghai: Shanghai education press, 2005
[26] Chang, Chih-Chung, and Chih-Jen Lin, "LIBSVM: A library for support vector machines." *ACM transactions on intelligent systems and technology (TIST)* vol. 2.3, No. 27, April. 2011.
[27] Krebel, UH-G, "Pairwise classification and support vector machines." *Advances in kernel methods: support vector learning*, pp. 255-268, 1999.
[28] Xu, Yi, "Contextual tonal variations in Mandarin." *Journal of phonetics*, vol. 25.1, pp. 61-83, 1997.
[29] Gauthier, Bruno, Rushen Shi, and Yi Xu, "Learning phonetic categories by tracking movements." *Cognition*, vol. 103.1, pp. 80-106, 2007.
[30] Peng, Gang, and William S-Y. Wang, "Tone recognition of continuous Cantonese speech based on support vector machines." *Speech Communication*, vol. 45.1, pp. 49-62, 2005.
[31] Chen, Sim-Horng, and Yih-Ru Wang, "Tone recognition of continuous Mandarin speech based on neural networks." *IEEE Transactions on speech and audio processing*, vol. 3.2, pp. 146-150, 1995.
[32] Lin, Ju, et al, "Improving Mandarin Tone Recognition Based on DNN by Combining Acoustic and Articulatory Features Using Extended Recognition Networks." *Journal of Signal Processing Systems*, vol. 90.7, pp. 1077-1087, 2018.
[33] Zhang J. *Foundation of Chinese Man-Machine Speech Communication*, Shanghai: Shanghai Scientific and Technical Publishers, 2010.
[34] Barras, Claude, and J-L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03)*, Hongkong, 2003, pp. II-49.
[35] Rose, Phil, "Considerations in the normalisation of the fundamental frequency of linguistic tone," *Speech communication*, vol. 6.4, pp. 343-352, 1987.
[36] Hombert, Jean-Marie, "Consonant types, vowel quality, and tone," *Tone: A linguistic survey*, pp. 77-111, 1978.
[37] Lin, Maocan, and Zhiqiang Li, "Focus and Boundary in Chinese Intonation," in *ICPhS*, Hongkong, 2011, vol. 17, pp. 1246-1249.
[38]