Speech Communication 168 (2025) 103173

Contents lists available at ScienceDirect

Speech Communication

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

Sequential perception of tone and focus in parallel–A computational simulation

Yue Chen *, Yi Xu

Department of Speech, Hearing and Phonetic Sciences, University College London, London WC1N 1PF, UK

ARTICLE INFO

Keywords:
Tone perception
Focus perception
Computational modeling
Gate recurrent unit (GRU)
Sequential speech processing
Co-current processing

ABSTRACT

Speech is produced continuously over time. So, the information it conveys, including intonational functions, also unfolds over time. But many intonational functions are encoded across whole utterances rather than only within certain words. How can perception process speech signals continuously over time, even for communicative functions that are globally encoded? In this study we used computational simulation to test the idea that even for intonational functions with large temporal scopes, it is possible to process f_0 contours syllable-by-syllable, and recognize the functions by continuous estimation of progressive probabilistic inference. We trained SVM and GRU models to simulate the perception of Mandarin tone and sentence focus with either syllable-sized or sentence-sized f_0 contours as input. The sentence-wide f_0 contours are gated at different syllable locations to test the incrementality of the recognition of tone and intonation. We also tested human listeners' perception of tone and focus with full and fragmented f_0 contours from the same dataset to evaluate the validity of the simulated perception. The results showed that the simulated syllable-by-syllable processing of tone and focus generated the closest recognition patterns to human perception. The simulations also show that there is little difference whether tone and focus are recognized separately or as tone-focus combinations, which suggests that despite sharing the same acoustic dimension, the two functions are sufficiently separated from each other in their f_0 coding.

1. Introduction

Perception is a critical phase in speech communication through which listeners decode continuous acoustic signals into discrete phonetic units. Because articulation proceeds sequentially, speech has to be received by the auditory system continuously as well. However, information conveyed by the speech signal is both multifaceted and distributed across different temporal scopes such as words, phrases and sentences. How can the multiple layers of information be decoded over time in perception? Is it done fully continuously or in steps with some kind of time window? Are the different layers of information decoded simultaneously, or separately in different time steps? And for units with a large temporal scope, does perceptual decoding have to rely on the pattern matching of the entire scope? Or identification proceeds in real time as the utterance unfolds? The present study tries to answer these questions by using computational modelling to simulate perceptual processing of tone and focus, two linguistic functions with very different temporal scopes. The language explored is Mandarin, which has been examined extensively not only for its tones (Peng and Zhang, 2015; Zhu and Wang, 2015), but also for the prosodic marking of focus (Chen, 2022; Xu, 2015).

1.1. Tone in Mandarin

Tones are pitch patterns that can distinguish words or grammatical functions in tonal languages (Yip, 2002), which are functionally analogous to consonants and vowels. Mandarin has four full lexical tones[1]: Tone 1 (high-level), Tone 2 (mid-rising), Tone 3 (low-dipping) and Tone 4 (high-falling) (Chao, 1968; Yip, 2002), whose fundamental frequency (f_0) contours in the syllable /ma/ spoken in isolation are shown Fig. 1.

When produced in connected speech, however, these tones exhibit rather different f_0 contours from those in Fig. 1. Some changes are due to drastic changes of tonal targets, e.g., Tone 3 loses its final rise when followed by any other tone, and changes into Tone 2 (or exhibits a contour very similar to Tone 2) when followed by another Tone 3 (Chao, 1968; Zhang, 2022). But most of the changes are due to speakers'

* Corresponding author.
E-mail address: yue.chen.1@ucl.ac.uk (Y. Chen).

[1] There is also a fifth tone in Mandarin known as the neutral tone, which is not included in the present study.

https://doi.org/10.1016/j.specom.2024.103173
Received 23 May 2024; Received in revised form 19 December 2024; Accepted 22 December 2024

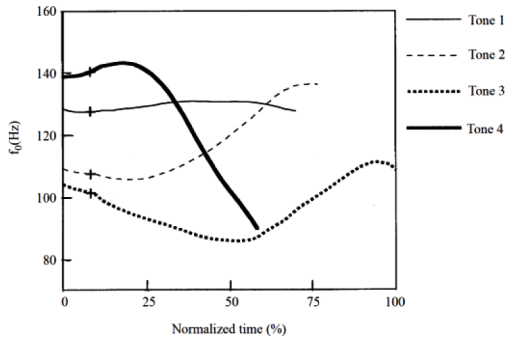

Fig. 1. Mean f_0 contours of four Mandarin tones in the monosyllable /ma/ produced in isolation (Xu, 1997).

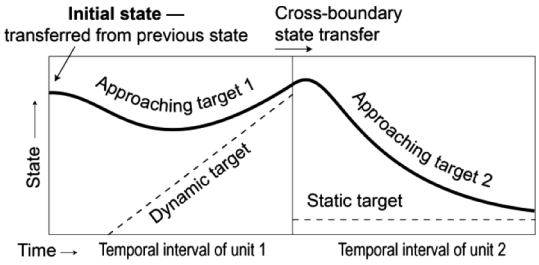

Fig. 2. The target approximation model (adapted from Xu and Wang, 2001). The dashed lines represent the underlying targets. The solid lines represent the f_0 realization. It demonstrates a dynamic tone (rising tone) followed by a static tone (low-level tone).

realization the underlying tonal targets under the combined constraint of inertia and tone-syllable synchronization (Xu, 2020; Xu and Wang, 2001), a mechanism characterized by the target approximation model, as illustrated in Fig. 2.

1.2. Prosodic focus

Focus is a communicative function to emphasize particular words within an utterance. Prosodic focus, the prosodic marking of emphasis, is a robustly encoded melodic function in Mandarin, whose phonetic realization is well documented (Chen and Braun, 2006; Jin, 1996; Shih, 1988; Xu, 1999). The encoding of prosodic focus in Mandarin involves a tri-zone pitch range modification: on-focus pitch range expansion, post-focus pitch range compression, and minimal or inconsistent pre-focus pitch range modification (Wang et al., 2018; Wang and Xu, 2011; Xu, 1999; Xu et al., 2012), which is shared with many non-tone languages (Alzaidi et al., 2019; Ardali and Xu, 2012; Bruce, 1982; Chahal, 2003; Dohen and Loevenbruck, 2004; Féry and Kügler, 2008; Ipek, 2011; Ishihara, 2003; Lee and Xu, 2010; Mixdorff, 2004; Patil et al., 2008, 2008; Rump and Collier, 1996; Wang et al., 2011). On-focus pitch range expansion results in more exaggerated underlying tonal targets: high pitches becoming even higher, and lower pitches even lower, as can be seen in Fig. 3 (Xu, 1999). Post-focus compression (PFC) results in both narrowing and lowering of the pitch ranges of all tonal contours, as can be also seen in Fig. 3.

1.3. Perceptual decoding of tone and focus–how is it done?

1.3.1. Perceptual cues for tone and focus

The fact that tone and focus both use f_0 as the main encoding property simultaneously thus raises serious questions for speech perception: How can they be teased apart from each other during perceptual processing? And, how can they be differentially processed given that very different temporal scopes are involved in their respective encoding? There have been various studies looking into the perception of tone and focus, respectively. But they have mostly examined the perception of the two types of functions separately, so the findings are not directly informative about how two overlaid functions can be perceived at the same time.

For tone, most perceptual studies have focused on establishing the most critical cues for tone identification. The different cues refer either to different acoustic properties such as fundamental frequency, amplitude, duration, and phonation (Blicher et al., 1990; Wang, 1972; Whalen and Xu, 1992; Yu and Lam, 2014), or to different dimensions of the f_0 contours, such as pitch height, pitch slope, pitch onset, etc. (Abramson, 1978; Gandour, 1983; Massaro et al., 1985; Shen and Lin, 1991; Wang, 1967). There is already much consensus that f_0 is by far the most important acoustic property of Mandarin tone perception (Howie, 1976;

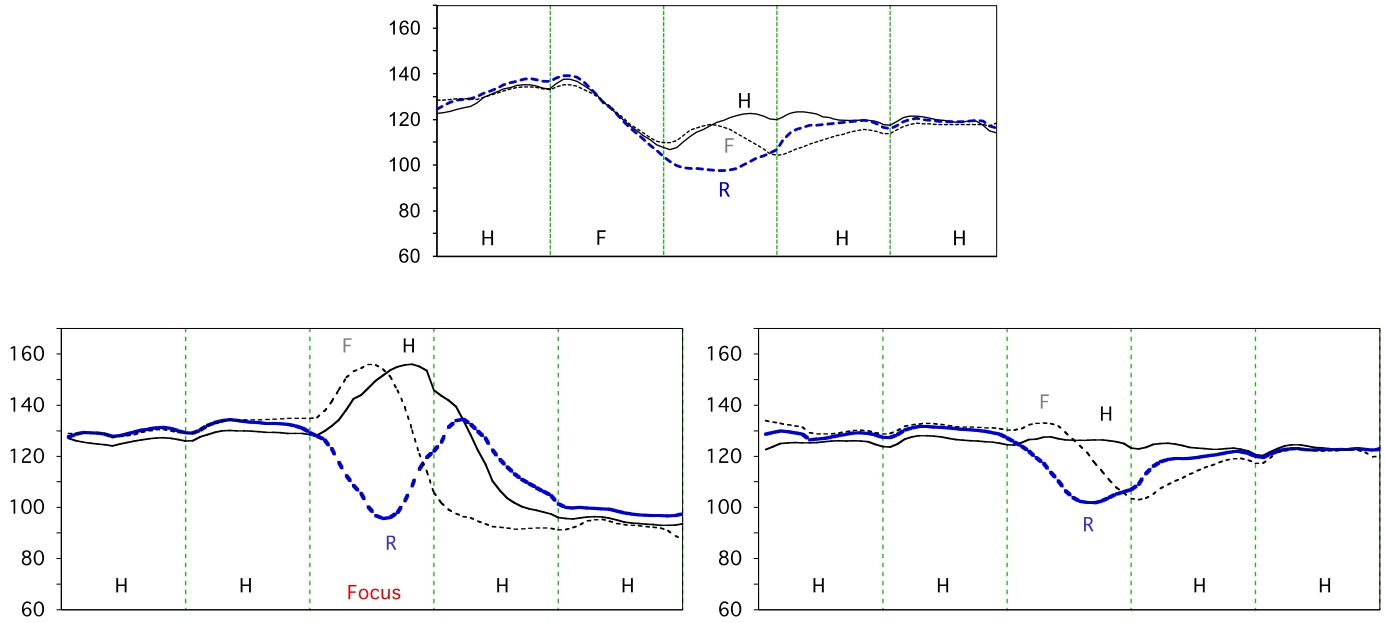

Fig. 3. Tri-zone encoding of focus in Mandarin as demonstrated by examples of focus on the third syllable (lower left plot) and focus on the first two syllables (lower right) as compared to neutral focus (top plot) (data from Xu, 1999). F, H, R represent falling tone, high tone and rising tone, respectively.

Liu and Samuel, 2004; Xu, 1997), but little agreement is reached as to which particular dimensions of f_0 contours are the most critical. Later studies tend to explore the weights of different cues for perception, assuming all main cues are used. Both the classical and more recent approaches, however, are based on the assumption that listeners first identify those cues and then weigh their relative importance for perception (Chandrasekaran et al., 2010; Francis et al., 2008; Leung and Wang, 2020; Tong et al., 2015; Tupper et al., 2020; Zhang et al., 2022). A very different approach was proposed by Chen et al. (2022), which raises the possibility that tone perception is done without pre-extracting cues or features. Instead, perception could simply process the whole f_0 contours of each syllable without isolating any specific cues before recognizing the tone category. This idea was tested with computational models that simulate both cue extraction and holistic processing, and the results showed that the latter not only had the best tone recognition rate, but also involved the lowest level of computational complexity.

For focus perception, many early studies have observed cues on the focused item only, often under the name of emphasis (O'Shaughnessy, 1979), sentence stress (Lehiste, 1970; Van Heuven, 2018), or pitch accent (Pierrehumbert, 1980; Silverman and Pierrehumbert, 1990). A focused item is found to exhibit expanded pitch range, longer duration, greater intensity, and possibly increased high-frequency spectral energy (De Jong, 2004; Sluijter and van Heuven, 1996). But later studies have shown that perceptual cues for focus are distributed across the full length of the sentence rather than located only at the position of focus. Rump and Collier (1996) find that the identification of focus type is jointly dependent on the f_0 of both the early and late target words in resynthesized short sentences in Dutch. Similar results were found in Mixdorff (2004) for Finnish in which the Fujisaki model was used to manipulate f_0 height of the early and late target words. Furthermore, non-sentence-final focus is better identified than final focus, because only the former allows PFC to be manifested (Botinis et al., 1999; Ipek, 2011; Lee et al., 2016; Liu and Xu, 2005). Also focus is perceptually more robust in languages that mark focus with PFC than languages without this character (Chen et al., 2009; Lee et al., 2015; Xu et al., 2012). The perception of focus therefore seems to rely heavily on the pitch of the post-focus syllables as well as the pitch of the focused word.

The importance of perceptual cues provided by sentence-wide f_0 profiles can be also seen from studies that ask listeners to identify focus from fragmented sentences. Botinis et al.'s (1999) show that words extracted from their original sentence context in English, Greek and Swedish have much lowered identification rate for their focus status. Xu et al. (2004) did a perception study with various parts of a sentence removed by replacing them with noise. They found that pitch of both on-focus and post-focus words provide critical information for focus perception. More specifically, focus could be recognized fairly well when either on-focus or post-focus portion was replaced by noise, and it could be recognized with high consistency when both on-focus and post-focus words were present. When neither on-focus nor post-focus words were available, it is quite difficult to identify the focus type, indicating that pre-focus words carry little focus cues.

1.3.2. Online perception of tone and focus: holistic vs. syllable-by-syllable

The finding that the focus encoding is global across the whole utterance raises a further question, namely, how exactly can focus be perceived together with tone which is largely local to individual syllables? One possibility is that listeners perceive focus by processing f_0 trajectories across the whole sentence regardless of local tones. This has been tested for Mandarin in Gauthier et al. (2009) using self-organizing map (SOM), and the results showed that the different focus positions could be effectively clustered from sentential pitch contours.

An alternative, however, is a sequential prediction process. That is, perception sequentially processes syllable-sized f_0 contours that convey not only tone information, but also focus information. In this way, listeners can parse focal intonation through local syllabic f_0 contours. Upon hearing each syllable, they not only categorize the lexical tone, but also predict the focus of the whole sentence. The focus prediction, however, remains partial until the whole sentence is heard. As found in gating experiments, listeners are able to make judgment of global intonation based on partial information carried by excised partial utterances (Face, 2005, 2007; Thorsen, 1980; Xu et al., 2004), but assign different weights to different parts of the utterances corresponding to certain functions (Face, 2007; van Heuven and Haan, 2002). Those findings indicate a potential sequential processing of intonation. What is unclear, however, is how listeners can make use of partial information to make global decisions and how this can be done concurrently with the perception of local tones, especially when there are simultaneous contextual tonal variations as well as cross-speaker differences.

1.3.3. Lessons from computational modeling of tone and intonation

Some lessons can be learned from computational modeling of the production of tone and intonation. The PENTA model, for example, assumes that lexical tone and intonational functions are encoded in parallel by jointly shaping syllabic pitch targets as articulatory goals (Xu, 2005). Surface f_0 contours are then generated by approaching successive underlying pitch targets through target approximation, as shown in Fig. 2. This has been tested with PENTAtrainer, a modeling tool based on PENTA (Xu and Prom-on, 2014), which can be trained with functionally annotated speech data to predict f_0 contours that can be checked against unseen utterances and evaluated by native listeners. It has been shown that focus, tone and word-level prosody generated this way are perceptually intelligible and natural sounding to native listeners of Mandarin (Xu and Prom-on, 2014) and Emirati Arabic (Alzaidi et al., 2023).

The finding that f_0 contours carrying cues for both focus and tone/word-level prosody information can be computationally generated with multi-functional syllable-sized pitch targets may suggest that listeners can also parse focal intonation through local syllabic f_0 contours. This possibility can already be seen in the findings of the gating experiments mentioned earlier, but those studies did not examine the joint perception of focus and lexical tone or word-level stress. Also, behavior studies alone cannot tell us how listeners overcome difficulties like variability due to context effects, speaker differences (Zhang et al., 2018; Zhang and Chen, 2016) and inter-functional interactions (Chen and Gussenhoven, 2008; Shen, 1989; Wang et al., 2020). But such mechanistic details could be explored by computational modeling work, because fine-structured model training could simulate the data-driven learning process that listeners have to go through when acquiring their language (Kuhl, 2004; Ullas et al., 2022; Werker and Yeung, 2005), and when maintaining their language skill in daily communication, which keeps refreshing and reshaping the perception skills after the initial acquisition. The perceptual mechanism could then be deduced to some extent through the explicit implementation of computational models.

The present study is an attempt to explore whether listeners can perceive focus and tone through syllable-sized f_0 contours using both computational modeling and human perception tests. The kind of modeling implemented is what we would like to call acoustic-functional front-to-end modeling, in which computational models are developed to perform real-life like tasks, with raw speech signal as input and speech category (tone and focus in the present study) as output that can be directly compared to human performance. This differs from the more commonly seen characterization-oriented modeling that aim to capture characteristics of human behavior but short of performing real-life-like tasks, such as TRACE model for spoken word recognition (McClelland, 2013; McClelland and Elman, 1986) which did not take continuous speech signals but abstracted phonological features as input. The front-to-end modeling is much harder to do, as it is more demanding than characterization-oriented modeling. In the case of perception, nevertheless, there have actually been plenty of successful front-to-end models, that is, automatic speech recognition systems developed for various purposes. For tone recognition, Zhang and Hirose (2004)

developed a tone-nucleus model, which recognizes tones by extracting f_0 contours from only the nuclear portion of a syllable to circumvent the effects of f_0 transitions between adjacent tones. Qian et al. (2007) proposed bi-tone and tri-tone units to model contextual effects of Cantonese tone recognition, which also avoided processing f_0 contours in the initial portion of each syllable. Lin et al. (2016, 2018) combined f_0 and segmental features in tone recognition, which improved recognition rate. Yu (2017) explored the role of temporal resolution for tone recognition and found that it had limited effect on Cantonese tone classification. Gogoi et al. (2020) used six f_0 features to train Mizo tone classifiers with both support vector machines (SVMs) and deep neural network (DNN). Yan et al. (2023) trained Mandarin tone recognizers with random forest, and found that feature fusion and optimization can simplify the algorithm of tone recognition on a monosyllable corpus. Those modelling achieved varying degrees of success, few of these studies, however, have made direct comparisons with human performance.

For intonation recognition, a number of studies followed the framework of auto-segmental metrical and developed automatic tools for ToBI, a speech prosody annotation system (Silverman et al., 1992). Rosenberg (2010) and Rosenberg et al. (2015), for example, developed computational tools that can generate ToBI annotations, including those of pitch accents, whose definition partially overlaps with prosodic focus (Ladd, 2008; Pierrehumbert and Hirschberg, 1990). Hu et al. (2020) introduced a similar system for Dutch called AuToDI. These systems first detect prominent syllables or words as pitch accents, and then classify them according to the shape and alignment of their pitch contours. There are also many attempts to explore the most efficient acoustic features and appropriate contextual information for word prominence/pitch accent detection (and/or prosodic boundaries) with different machine learning models (Ananthakrishnan and Narayanan, 2005; Fernandez and Ramabhadran, 2010; Jeon and Yang Liu, 2009; Kakouros et al., 2018, 2019; Kakouros and Räsänen, 2016; Levow, 2005; Mishra et al., 2012; Ren et al., 2004; Schnall and Heckmann, 2019; Stehwien and Vu, 2017; Walsh et al., 2013). All these works have aimed at improving the accuracy of pitch accent classification.

None of the automatic prosody recognition systems reviewed above, however, have been designed to simultaneously process lexical tone and focus. Also, with tone or prominence recognition as the sole objective, they are not concerned with theoretical questions about the perception of tone and intonation. Additionally, for the sake of maximizing performance, as many input features as possible are included in these models, regardless of which features are critical for perception and which are not. For answering theoretical questions about simultaneous perception of tone and focus, therefore, new task-driven models need to be developed. These models should a) be built to perform front-to-end recognition tasks for both tone and focus, as opposed to only generating characterizations of perception patterns, b) be able to recognize tone and focus from unfinished utterances to simulate progressive perception, c) combine the data-driven modelling and theoretical rules (theory enhanced data-driven modelling) and d) be validated with human behavior data.

These front-to-end models will not be built to recognize tone and focus through cues such as f_0 height, f_0 slope, or descriptive f_0 profiles, etc., as they have already been found to be less effective than raw f_0 contours (Chen et al., 2022). Also, the models will process only f_0 data, without other data such as duration, intensity, voice quality and spectral properties, etc., as it is already found that f_0 carries sufficient information for both tone and focus in production (Prom-on et al., 2009; Xu and Prom-on, 2014) as well as perception (Alzaidi et al., 2023; Mixdorff, 2004; Rump and Collier, 1996), although other information may also help (Prom-on et al., 2009). In this study, we want to explore the power of modeling in a most straightforward scenario possible. The effectiveness of the modeling will be checked against human perception to find out the model fitting human performance the best and how much information is missing without the non-f_0 cues.

Table 1
Sentences used as recording materials and their tone patterns. H, R, L, and F represent high, rising, low, and falling tones, respectively (Xu, 1999).

Word 1	Word 2	Word 3
HH 猫咪/māomī/ 'Kitty'	H 摸/mō/ 'touches'	HH 猫咪/māomī/ 'kitty'
HR 猫迷/māomí/ 'Cat-fan'	R 拿/ná/ 'takes'	LH 马刀/mǎdāo/ 'sabre'
HL 猫米/māomǐ/ 'Cat-rice'	F 卖/mài/ 'sells'	
HF 猫蜜/māomì/ 'Cat-honey'		

1.4. Current study

For computational modeling to be relevant for enhancing our understanding of speech perception, it should allow us to explore the feasibility of various conceivable perceptual strategies. In a pilot study, we tested the feasibility of modeling parallel recognition of tone and focus in Mandarin by processing syllable-sized local pitch targets (Chen and Xu, 2021). The current study is to extend this work by building models that can simulate different possible strategies of perceiving tone and focus. In the time domain, the key question is whether focus can be processed syllable-by-syllable although its full temporal scope covers the whole sentence (Xu et al., 2004). In the functional domain, the crucial issue is whether tone and focus are processed hierarchically or independently. We therefore tried to answer the following research questions through computational simulation of tone and focus perception on a Mandarin corpus.

1. Can tone and focus be recognized independently of each other, or they have to be co-processed for concurrent recognition?
2. Does focus have to be recognized by processing sentence-wide f_0 contour as a whole, or the recognition can be done syllable-by-syllable, guided by progressive accumulation of probability?

To answer these questions, we applied support vector machine (SVM), a non-neural network model, and Gated Recurrent Unit (GRU), a recurrent neural network model, for acoustic-phonetic learning. In sequential focus processing simulations, we utilized Bayesian inference to integrate sub-functional information into global focus categories.

As will be introduced in the following section, a set of modelling experiments were carried out to assess various perception strategies with full or fragmented utterances tested in terms of recognition accuracy and confusion patterns. The primary goal of the present study is to find out the most plausible perception mechanism by comparing the results of human perception and model recognition of Mandarin tones and focus.

2. Materials and methods

The overall strategy is to develop computational models that are trainable with f_0 contours from connected speech to recognize tone and focus. The ability of these models to simulate human perception is tested by comparing model performance with human perception of tone and sentence focus. The computational models are configured differently to simulate several alternative perception strategies, and the efficacy of each strategy is estimated in terms of recognition outcomes compared with human perception patterns.

2.1. Corpus and annotation

The corpus used in this study is an experimental Mandarin dataset collected in Xu (1999). All the sentences consist of five syllables (three words) with varying tones on the middle three syllables (see Table 1), and the first word (subject) and the third word (object) are disyllabic and the second (verb) is monosyllabic. The first, second, and third words have four, three and two alternatives, respectively. Thus, there are 4 (1st word) × 3 (2nd word) × 2 (3rd word) = 24 target declarative sentences

Table 2
F_Syl1—Focus labeling Scheme 1 for marking focus status of each syllable corresponding to each sentential focus category.

Focus	1st syllable	2nd syllable	3rd syllable	4th syllable	5th syllable
Neutral	Neutral/ Pre-	Neutral/ Pre-	Neutral/ Pre-	Neutral/ Pre-	Neutral/ Pre-
Initial	On-	On-	Post-	Post-	Post-
Middle	Neutral/ Pre-	Neutral/ Pre-	On-	Post-	Post-
Final	Neutral/ Pre-	Neutral/ Pre-	Neutral/ Pre-	On-	On-

Table 3
F_Syl2—Focus labeling Scheme 2 for marking focus status of each syllable corresponding to each sentential focus category.

Focus	1st syllable	2nd syllable	3rd syllable	4th syllable	5th syllable
Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
Initial	On-	On-	Post-	Post-	Post-
Middle	Pre-	Pre-	On-	Post-	Post-
Final	Pre-	Pre-	Pre-	Final-on-	Final-on-

training, and 1 repetition for evaluating the trained model.

2.2. Data

The raw source data were continuous sentence-sized f_0 contours with syllable boundaries annotated, as illustrated in Fig. 4. Each syllable was represented by a 10 data point vector taken from the time-normalized syllable-sized f_0 values in both Hertz and semitones (with 1 Hz as the reference f_0). Also extracted were velocity profiles at 10 time-normalized points per syllable. All those f_0 profiles were extracted using ProsodyPro (Xu, 2013). In addition, to address the potential effect of variations across speakers as well as repetitions, a new f_0 profile, Δf_0, was computed which is the difference between the f_0 value of current point and the onset f_0 of the sentence.[2] In total, 5 input features were extracted: f_0 in Hz (f_0), Δf_0 in Hz (Δf_0), f_0 in Semitone (Semitone), Δf_0 in Semitone (ΔSemitone) and f_0 velocity (Velocity) for respective modelling. No other data normalization or pre-processing was applied.

2.3. Computational models

In a previous study (Chen et al., 2022), we have already found that SVM can achieve good tone recognition from raw f_0 contours. So, SVM is again used in this study as one of the models to further test its power on focus recognition. However, neural networks nowadays have shown their strong ability in language processing. Recent studies compared brain responses of human and the processing activity of a neural network and found that artificial intelligence (AI) systems can process signals in a way that is similar to how the brain interprets speech sound (Beguš et al., 2023; Li et al., 2023). So, we also want to try neural network models on tone and focus recognition to see if better performance than SVM can be achieved. All the models used in this study are supervised rather than unsupervised (e.g., Gauthier et al. 2007, 2009). Unsupervised models would simulate learning in which the learner discovers linguistic categories from speech signals without knowledge of phonetic categories. Phonetic acquisition research, however, has shown evidence that knowledge of phonetic category is acquired through social interaction and is used to guide the development of language-specific phonetic perception (Kuhl, 2010; Kuhl et al., 2014). So, despite findings that unsupervised phonetic learning is possible (Gauthier et al., 2007, 2009), only supervised learning models are applied in the current study.

2.3.1. Gated recurrent unit (GRU): a recurrent neural network

Given that speech is generated continuously, a recurrent neural

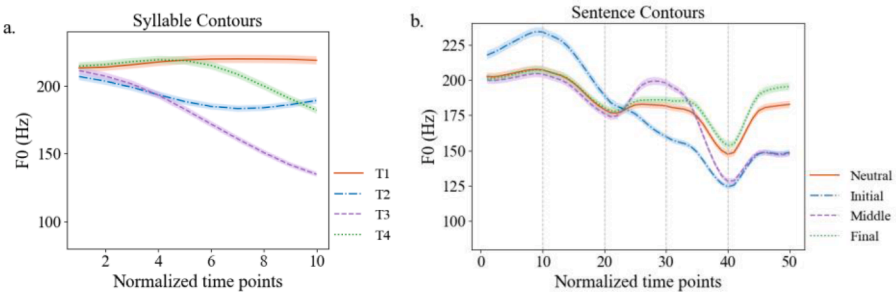

Fig. 4. Plots of average f_0 contours (lines) with standard error (shade). Left: f_0 contours of the second syllable in all sentences. Right: f_0 contours of whole sentences under different focus conditions.

in the corpus. There are four citation tones in Mandarin, including Tone 1 (high), Tone 2 (rising), Tone 3 (low), and Tone 4 (falling). No neutral tone or cases of tone sandhi were included in the corpus. The sentences were recorded by four male and four female native Standard Chinese (Putonghua) speakers. Each sentence was spoken with four different focus patterns: focus on the first (initial), second (middle), or third word (final), and neutral focus, elicited by WH-questions. In total, there are 3840 sample sentences (24 basic sentences × 4 focus patterns × 8 speakers × 5 repetitions).

The corpus was annotated with syllable boundaries, lexical tones and focus. The annotation of syllable-level focus events was done in two different ways to represent two alternative hypothetical focus encoding schemes. Focus Scheme 1 (F_Syl1), as shown in Table 2, consists of only three different focus labels for each syllable, which treats all words in a neutral-focus sentence as the same as pre-focus words in sentences with non-initial focus. This was based on a strict interpretation of the tri-zone hypothesis of focus (Xu, 2005; Xu et al., 2004). Focus Scheme 2, as shown in Table 3, consists of five focus labels, which annotates all syllables in neutral focus sentences as Neutral and syllables in the last word in a final-focus sentence as Final-on. The separate annotation of final-focused syllables is based on findings that final focus is less robustly encoded than non-final focus in Mandarin (Xu et al., 2012) as well as in many other languages (Botinis et al., 1999).

The whole dataset was divided into a training subset, a validation subset and a testing subset, with a ratio of 3:1:1, whereby 3 random repetitions of each sentence by one speaker were used for training the recognition model, 1 repetition for optimizing the model during the

[2] The use of such Δf_0 has been found to be effective in simulating the learning of tone and intonation production (Meng et al., 2023; Prom-on et al., 2009; Xu and Prom-on, 2014).

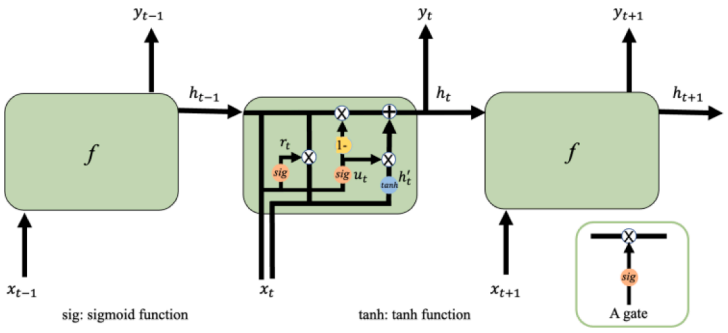

Fig. 5. A simple demonstration of a GRU layer with three cells.

network (RNN) may be an appropriate model for simulating speech perception, especially for focus that is encoded across a whole sentence as discussed earlier. Here, we chose Gated Recurrent Unit (GRU), a special type of RNN proposed by Cho et al. (2014).

RNNs are the neural networks where the hidden state from the previous step is fed as input to the current step so as to process sequential data. To solve the problem of gradient vanishing or exploding due to long-term dependency in vanilla RNN, some gating mechanisms are developed to selectively update the hidden state at each time step to be sent to next step. Compared with the more widely used long-short term memory (LSTM), GRU only have two gates in a cell which involves less parameters and faster speed to compute. Fig. 5 gives an example structure of a GRU layer.

The simulation process was as follows:

$$r_t = sigmoid(w^r[h_{t-1}, x_t] + b^r)$$

$$u_t = sigmoid(w^u[h_{t-1}, x_t] + b^u)$$

$$h'_t = tanh\left(w^h[r_t * h_{t-1}, x_t] + b^h\right)$$

$$h_t = (1 - u_t) * h_{t-1} + u_t * h'_t$$

The data processing within a GRU cell can be seen as a function: $h_t = GRU(h_{t-1}, x_t)$. The subscript t represents the time step. The input of each GRU cell (time step) has two sources: the present input (x_t) and the preceding hidden state (h_{t-1}). The output at each time step is the copy of hidden state h_t. Looking into the cell, each cell has two gates: reset gate r_t and update gate u_t. The reset gate r_t decides how much of the information should be kept from h_{t-1}, which is then used to generate the candidate hidden state h'_t reducing the effect that previous information has on the current information. The update gate u_t decides how much information to forget from h_{t-1} and how much information to add from h'_t to update the current hidden state h_t. w^r, w^u and w^h are learnable weight matrices, and b^r, b^r and b^r are the bias terms. After the GRU layers, there will be a fully connected layer. The output of the GRU layer will be put into a fully connected layer to generate the probabilistic class predictions through a softmax function, which are tailored to the specific requirements of the recognition task.

A bidirectional GRU (Bi-GRU) is a GRU neural network which does not change the inside structure of the cell but runs in two directions within a (Bi-)GRU layer. The results of the two processing in a Bi-GRU layer are combined to generate the output of the layer, which means both the previous and the following time steps can affect the current time step. The output of a Bi-GRU cell is: $h_t = [GRU(h_{t-1}, x_t); GRU(h_{t+1}, x_t)]$.

In this study, we trained both GRU and Bi-GRU with Pytorch (Paszke et al., 2017) for tone and focus recognition. The training set was used to train the models and the validation set was used to optimize the models. The testing set was put into the trained models to estimate the effectiveness of different strategies. The detailed procedures will be introduced later in the Section 2.4.

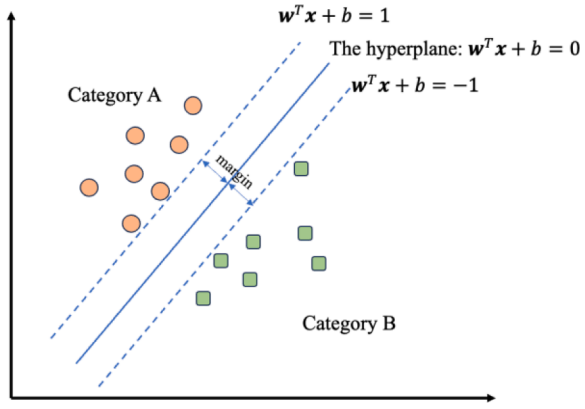

Fig. 6. A simple demonstration of a linear SVM.

2.3.2. Support vector machine (SVM)

SVM is a supervised machine learning model widely used in speech recognition and can handle both linear and non-linear classification tasks. SVM was originally developed for binary classification tasks, and the basic idea of SVM is to represent samples as points in a space and find a clear hyperplane or gap that is as wide as possible to separate categories. Fig. 6 gives a simple example of binary linear SVM classifier. The hyperplane can be defined as:

$$\boldsymbol{w}^T\boldsymbol{x} + b = 0$$

where $\boldsymbol{w}$ is the weight vector, $\boldsymbol{x}$ is the input vector, and b is the bias term. Before training, all the samples are labelled as $+1$ or -1. If the label is $+1$, $\boldsymbol{w}^T\boldsymbol{x} + b$ is expected to be larger than $+1$, otherwise it is smaller than -1. The data points that have the smallest perpendicular distance to the hyperplane are called support vectors (on the dashed lines). The hyperplane is actually determined by those vectors. The weight $\boldsymbol{w}$ shows how each dimension of those vectors is used in the classification process. The larger the margin, the better the hyperplane. The test samples are then mapped into the same space and classified based on which side of the gap they fall.

In this study, the SVMs were trained using the Scikit-learn tool (Pedregosa et al., 2011) with RBF (Radial Basis Function) kernel to transform the input data into higher-dimensional space to enable linear separation. For multiclass SVM classifier, we adopted the OVO (One-vs-One) strategy generalizing 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 (Kreßel, 1999). Grid search and cross-validation were used to tune the hyperparameters: regularization (c), and gamma values (g). The training set and validation set were combined for SVM training, and five-fold cross-validations were automatically and randomly applied during the training to optimize the model.

2.4. Computational simulations

The simulations were aimed at answering the two questions outlined in 1.4. The first question is whether tone and focus can be recognized independently of each other, or they have to be co-processed for concurrent recognition. The second question is whether focus has to be recognized by processing sentence-wide f_0 contour as a whole, or the recognition can be done syllable-by-syllable, guided by progressive probabilistic inference. Firstly, to test if focus can be recognized syllable-by-syllable or holistically, we set up two kinds of models. The syllable-by-syllable simulation recognizes syllable-level focus events and integrates the syllabic probabilities into sentential focus decisions. Given that coarticulation or contextual information may affect the performance of local recognition, we designed two sets of local recognizers. One of them recognized local event individually regardless of the context, and the other one recognized local event in a sequential context. The holistic simulation recognizes sentence level focus directly from whole-sentence f_0 profiles. Secondly, we wanted to investigate if focus and tone are recognized independently or hierarchically. To simplify the hierarchical correlation assumed, we designed a simultaneous recognition task of tone and focus through the combination of tone and focus categories, which can be compared with tone-only recognition and focus-only recognition.

To sum up, we designed four recognition tasks: 1) lexical tone recognition, 2) syllable-by-syllable focus recognition, 3) direct sentence-wide focus recognition and 4) simultaneous syllable-by-syllable tone and focus recognition. For each task, both neural network (GRU) and non-neural network model (SVM) were trained to achieve maximum recognition rate based on different input f_0 profiles (f_0, Δf_0, Semitone, ΔSemitone and Velocity), respectively. We used unweighted average recall (UAR), which is the average of the recall on each class, and confusion matrix, which is a summary of correct and incorrect predictions broken down by each class, to choose the model that fits the human performance the best, and analyzed the possible mechanisms the perceptual process may involve.

2.4.1. Experiment 1: recognition of tone only

In this experiment, all models performed only tone recognition tasks on syllable-sized f_0 contours. Each syllable was labelled only for tone: T1, T2, T3 and T4. In this way, the models were trained to recognize the tones *regardless of* the focus condition of the sentence.

Five models were built based on different underlying mechanisms:

1. T-SVM — A SVM model that recognizes lexical tones locally by classifying them based on Euclidean distance. For each sample, the model takes a 10 equidistant discrete point vector from a syllable as the input. The training target of each sample is the tone category and class membership probability estimates are enabled.
2. T-GRU — A GRU model that recognizes lexical tones locally through a unidirectional recurrent neural network. It takes one syllable as a time sequence and the sequence length is 10. The input at each time step is one f_0 point. The model contains two unidirectional GRU layers and one fully connected layer at the last time step with 128 hidden units in each hidden layer. A GELU activation function is applied to produce the probabilities of the four tones as output. During training, cross-entropy loss function and the Adam optimizer were used with dropout rate of 0.1 and batch size of 32.
3. T-Bi-GRU — A GRU model with the same main structure as T-GRU but the GRU layers are bidirectional. It assumes that lexical tones can be recognized locally through a bidirectional recurrent neural network.
4. T-GRU-Con — A GRU model that recognizes lexical tones through a neural network that also takes preceding f_0 context as part of the input. It takes the whole sentence as a sequence and the sequence length is 5. The input at each time step is a 10 equidistant discrete point vector from one syllable. The model also contains two unidirectional GRU layers and one fully connected layer with 128 hidden units in each hidden layer. A GELU activation function is applied to produce the probabilities of the four tones for output at each time step. During training, cross-entropy loss function and the Adam optimizer were used with dropout rate of 0.1 and batch size of 32.
5. T-Bi-GRU-Con — A GRU model with the same main structure as T-GRU-Con but the GRU layers are bidirectional. It assumes that lexical tones can be recognized through a neural network and affected by both preceding and following contexts.

Table 4
Sentence fragments of varying lengths used as stimuli.

Number of syllables	Sentence fragments
1	2nd syllable of Word 1
2	Word 1
3	Word 1 + Word 2
5	Word 1 + Word 2 + Word 3

For the last two models that take the global context into consideration, T-GRU-Con and T-Bi-GRU-Con, we also tested their recognition capabilities on fragments of the sentences with only the second syllable, the first (disyllabic) word and the first two words to see if contexts have significant effect on tone perception. Table 4 gives the four stimuli of sentence fragments for tone recognition. The fragments were also used in the subsequent recognition and perception tasks.

2.4.2. Experiment 2: syllable-by-syllable focus recognition

This experiment tests the recognition of focus without knowledge of tone in a syllable-by-syllable manner, in two steps: 1) recognizing local focus events; and 2) integrating local decisions into a global classification.

The first step is conducted by recognizing a sequence of syllable-level focus events. Similar to tone recognition, there are also five syllable-level focus recognition models:

1. F-SVM — A SVM model that recognizes syllable-level focus events locally by classifying them based on Euclidean distance.
2. F-GRU — A GRU model that recognizes syllable-level focus events locally through a unidirectional recurrent neural network.
3. F-Bi-GRU — A GRU model that recognizes syllable-level focus events locally through a bidirectional recurrent neural network.
4. F-GRU-Con — A GRU model that recognizes syllable-level focus events through a neural network that also takes preceding f_0 context as part of the input.
5. F-Bi-GRU-Con — A GRU model that recognizes syllable-level focus events through a neural network that also takes both preceding and following contexts as part of the input.

The model structures are the same as tone recognition models, respectively, but tasked to recognize two sets of syllable-level labels for focus (Tables 2 and 3): F_Syl1 and F_Syl2. F_Syl1 consists of three labels: Neutral/Pre-, On- and Post-. This labeling scheme assumes that all syllables in a neutral-focus sentence have the same focus status as pre-focus syllables. F_Syl2 consists of five labels: Neutral, Pre-, On-, Post- and Final-on-. This labeling scheme therefore treats syllables in neutral focus sentences differently from pre-focus syllables, and syllables under final focus differently from other on-focus syllables.

To achieve syllable-by-syllable focus recognition, the local recognition outcomes need to be converted to sentential decisions. This was done by applying Bayesian inference which is widely used in cognitive studies (Feldman et al., 2009; Kleinschmidt and Jaeger, 2015; Norris et al., 2016; Norris and McQueen, 2008) as follows.

The testing of the local recognition model generated a $1 * n$ matrix of probabilities for each syllable in the testing set and a $5 * n$ matrix ($\boldsymbol{P}_{5*n}$)

of probabilities for each sentence, where n is the number of categories of syllable-level focus events.

The local focus event label set is

$$\boldsymbol{SYL} = \begin{cases} [1 = \text{neutral/pre}, 2 = \text{on}, 3 = \text{post}], \ \textit{for F\_Syl1 and } n = 3 \\ [1 = \text{neutral}, 2 = \ \text{pre}, \ 3 = \text{on}, \ 4 = \text{post}, 5 = \ \text{final on}], \ \textit{for F\_Syl2 and } n = 5 \end{cases}$$

And the finite label set for the sentence is

$$\boldsymbol{SEN} = [1 = \text{neutral}, 2 = \text{initial}, \ 3 = \text{middle}, \ 4 = \text{final}] \in \mathbb{R}^4$$

Following Tables 2 and 3, each focus condition corresponds to one unique syllabic focus sequence. Hence the mapping matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{cases} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 3 & 3 & 3 \\ 1 & 1 & 2 & 3 & 3 \\ 1 & 1 & 1 & 2 & 2 \end{bmatrix}, \ \textit{for F\_Syl1} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 3 & 3 & 4 & 4 & 4 \\ 2 & 2 & 3 & 4 & 4 \\ 2 & 2 & 2 & 5 & 5 \end{bmatrix}, \ \textit{for F\_Syl2} \end{cases}$$

The value of element a_{ij} in $\boldsymbol{A}$ represents the local focus label for the j^{th} syllable in the sentence which belongs to the i^{th} category in $\boldsymbol{SEN}$.

The identification of the sentence-level focus category $c \in \boldsymbol{SEN}$ is to choose a syllabic label sequence $s_1, s_2, s_3, s_4, s_5 \in \boldsymbol{SYL}$ having a corresponding sentence focus category c that is the most probable given the observation sequence $O = \ o_1, \ o_2, \ o_3, \ o_4, \ o_5$. According to Bayesian inference:

$$\widehat{c} = \ argmax_c \ P(c|O) = argmax_c \frac{P(O|c)P(c)}{P(O)}$$

Because it is hard to decide the prior probability $P(c)$, we just assume that sentence focus probabilities are evenly distributed, as the numbers of the four focus conditions are equal in our corpus. $P(O)$ is a constant. Thus, we have:

$$\widehat{c} = argmax_c \ P(O|c) = argmax_c \prod_{t=1}^{5} P(o_t|s_t)$$

The likelihood probability $P(o_t|s_t^c)$ can be obtained from the syllable-level focus events recognizers above. Thus:

$$\widehat{c} = argmax_c \prod_{t=1}^{5} \boldsymbol{P}_{t,a_{ct}}$$

Then, the results of local focus events recognition can be converted to make the final decision of sentence focus category.

To examine whether sentence focus perception is syllable-by-syllable or holistic, not only do we need to test the models on the whole sentences, but also we have to see whether the models can simulate the incremental process that can be comparable to human perception performance. As the latter two models (F-GRU-Con and F-Bi-GRU-Con) take the global context into consideration, we have to test the trained models with f_0 contours not only from whole sentences f_0 contours, but also from the first two syllables (word 1) and the first three syllables (words 1 and 2).

As the training sample and testing sample should have the same data shape, the missing part of the sentence f_0 contours should be padded. Normally in RNN models, 0 is used to pad the sequence. However, although, in this task, 0 does not overlap with actual value of f_0 in this corpus, it is still a meaningful f_0 value and may make the sequences biased towards certain focus category. To minimize the bias, we used the grand average f_0 of the corpus instead of 0 as the padding value.

After syllabic focus recognition, sentential focus is predicted by the probabilities of the syllable from the fragmented sentence f_0 contours.

The sentential focus prediction of the first two words (the first three syllables) would be:

$$\widehat{c} = argmax_c \prod_{t=1}^{3} \boldsymbol{P}_{t,a_{ct}}$$

And the sentential focus prediction of the first word (the first two syllables) would be:

$$\widehat{c} = argmax_c \prod_{t=1}^{2} \boldsymbol{P}_{t,a_{ct}}$$

When making sentential focus prediction from incomplete sentence f_0 contours, a problem may arise that a sub-sequence of local focus events is shared by more than one sentence focus category. For example, with F_Syl1, neutral focus, middle focus and final focus sentences all have the same two-syllable sentence fragments for the first two syllables (both are neutral/pre). Once such sub-sequence obtained the highest accuracies, it would cause an identification ambiguity between those focus categories sharing the sub-sequence. Thus, we forced the program to take neutral focus as the final decision when neutral focus was one of the options or assumed that focused word would appear as early as possible when the sub-sequence was shared by middle focus and final focus.

2.4.3. Experiment 3: recognition of tone-focus combinations

Experiments 1 and 2 are for testing whether tone and focus can be recognized without knowledge of each other. Experiment 3, in contrast, examines whether it is beneficial to recognize them hierarchically, i.e., as syllable-sized tone-focus combinations. The task of the models is to sequentially recognize syllable-sized f_0 contour as unique tone-focus combinations that carry information of both functions. There are 12 labels for the combination of tone and F_Syl1, and 20 labels for tone and F_Syl2, which are used in the training of five recognition models:

1. TF-SVM.
2. TF-GRU.
3. TF-Bi-GRU.
4. TF-GRU-Con.
5. TF-Bi-GRU-Con.

The structures of the five models are the same as in tone only recognizers (Experiment 1) but are tasked to process two sets of tone-focus combination labels. The outcomes of the models are broken down into tone and focus events, respectively. The local focus events are then converted into sentential focus decisions through Bayesian inference. Those models are also tested on sentence fragments following the same routine as in Experiment 2.

2.4.4. Experiment 4: holistic focus recognition

This experiment tests sentence-wide holistic processing strategy of

Table 5
Tone patterns and corresponding sentences used in perception experiment.

Word 1	Word 2	Word 3
HH 猫咪/māomī/ 'Kitty' HR 猫迷/māomí/ 'Cat-fan' HL 猫米/māomǐ/ 'Cat-rice' HF 猫蜜/māomì/ 'Cat-honey'	H 摸/mō/ 'touches'	HH 猫咪/māomī/ 'kitty'

focus perception. The recognition models are trained on f_0 contours of whole sentences, but tested with both whole-sentence f_0 contours and fragmented f_0 contours. Like Experiment 2, focus recognition is done without knowledge of the tones in the sentences. Three different models were built for this task.

1. SF-SVM — A SVM model that recognizes focus based on Euclidean distance. For each sample, it takes a 50 (5 × 10) equidistant discrete point vector from a sentence as the input. The training target of each sample is the focus category.
2. SF-GRU — A GRU model that recognizes focus directly through a unidirectional recurrent neural network. It takes the whole sentence as a sequence and the sequence length is 5. The input at each time step is a 10 equidistant discrete point vector from one syllable. The model contains two unidirectional GRU layers and one fully connected layer at the last time step with 128 hidden units in each hidden layer. A GELU activation function is applied to produce the probabilities of the four focus categories for output. During training, cross-entropy loss function and the Adam optimizer were used with dropout rate of 0.1 and batch size of 32.
3. SF-Bi-GRU— A GRU model with the same main structure as SF-GRU but the GRU layers are bidirectional. It assumes that focus can be directly recognized through a bidirectional recurrent neural network.

The three holistic recognizers are also tested on the sentence fragments to simulate the incremental processing of focus.

2.5. Human perception experiment

To evaluate the performance of the computational recognizers, human listeners were asked to identify tone and focus from the same corpus.

2.5.1. Stimuli

Table 5 shows the sentences used in the perception experiment. Except syllable 2, all other syllables have the H tone (Tone 1). Only one of the five repetitions of each sentence was randomly chosen as stimulus to shorten the experiment. Thus, there were a total of 128 complete sentence stimuli = 4 base sentences * 4 focus conditions * 8 speakers * 1 repetition.

For testing progressive focus perception, some fragmented stimuli were also created containing only the second syllable, the first word, and the first two words, respectively. A total of 512 stimuli (128 sentences * 4 fragments) were therefore used.

2.5.2. Subjects

20 native Mandarin speakers, 10 females and 10 males, were recruited on Prolific (www.prolific.co). All of them were born and raised in China and aged between 23 and 40 years old. None reported any hearing impairment.

2.5.3. Procedure

The perception experiment was conducted on the Gorilla Experiment Builder (www.gorilla.sc) (Anwyl-Irvine et al., 2020). Listeners were given one sound at a time and were asked to choose the sentence with the correct tone and focus, as displayed on screen as in Fig. 7. The first selection is for the lexical tone, represented by different Chinese characters for the second syllable. The second selection is for the position of the focused word. The subject could hear each sound up to three times before making their two selections and moving to the next trial. Before the real trials, they were given 5 random trials (without feedback) to become familiar with the procedure.

3. Results

3.1. Experiment 1: tone recognition only

3.1.1. Overall accuracy

Five computational models were trained to recognize tones without knowledge of focus. As the number of syllables with the four tones are not equal in our corpus, we use the unweighted average recall (UAR) to evaluate the tone recognition models. Fig. 8 shows overall recognition accuracies (UARs) of the five models trained with five formats of f_0 data. Of the five models, three recognized tones without f_0 contexts (T-SVM, T-GRU and T-Bi-GRU), and two took f_0 context into consideration (T-GRU-Con and T-Bi-GRU-Con). The overall accuracies of the first three models are all above 85 %, and those of the last two are both over 95 %.

Fig. 9 shows UARs of the second syllable from T-GRU-Con and T-Bi-GRU-Con tested with fragmented f_0 contours of different lengths. With only f_0 contours of the second syllable, both models recognized the tones only at chance level (25 %). With f_0 of the first two syllables, the performances increased dramatically. With more f_0 contours after the

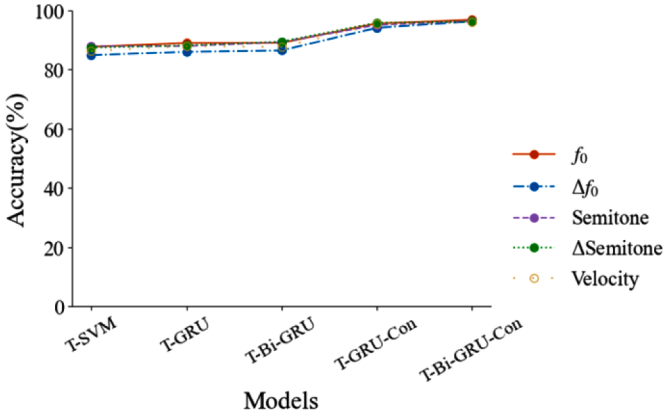

Fig. 8. Overall UAR of the five tone recognition models.

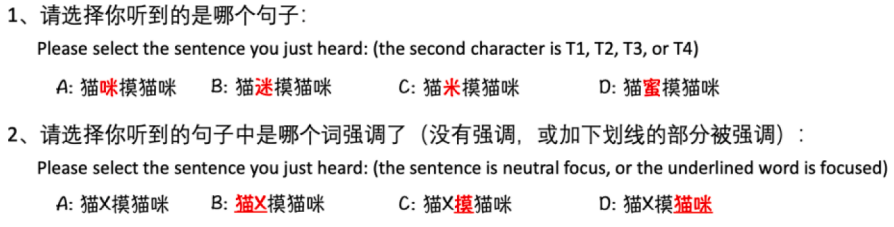

Fig. 7. Instructions and options in perception experiment.

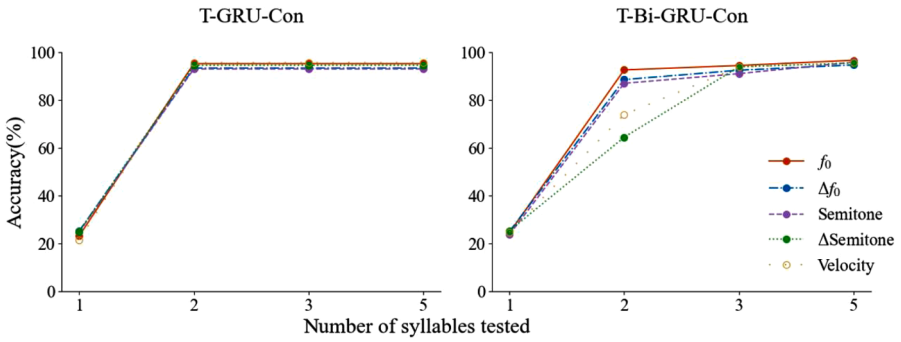

Fig. 9. Overall UARs of tone recognition as a function of the number of tested syllables from T-GRU-Sen (left) and T-Bi-GRU-Syl (right).

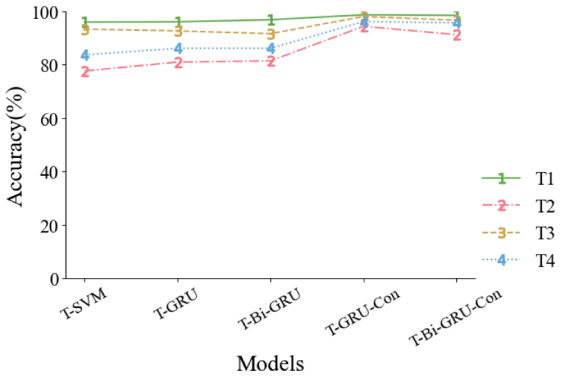

Fig. 10. Recognition rates of the four tones from the five tone recognition models tested with syllables of whole utterances.

second syllables included, the accuracy of T-GRU-Con (left) remained high at around 95 %, and the accuracy of T-Bi-GRU-Con (right) gradually rose from 93 % at two-syllable to 97 % at whole sentence when used f_0 in Hz as input.

3.1.2. Accuracy of individual tones

Figs. 8 and 9 both show that, for all the five models, there is not much difference between the five formats of f_0 data, but raw f_0 in Hz almost always has the highest accuracies. So, Fig. 10 shows the accuracies of the four tones with f_0 in Hz as input. For all models, the static tones always had higher accuracies than the dynamic tones: T1 > T3 > T4 > T2.

3.2. Experiment 2: syllable-by-syllable focus recognition

For syllable-by-syllable focus recognition without knowledge of tone, five models with the same structure as tone recognizers were trained to recognize syllable-level focus events based on two sets of labels as shown in Tables 2 and 3. The local recognition outcomes were converted to sentential focus through probabilistic inference.

3.2.1. Overall accuracy

Fig. 11 (a–e) shows overall accuracies of sentence focus obtained from the five syllable-by-syllable models. The left column shows results using the F_Syl1 labeling scheme and the right column shows results with the F_Syl2 scheme. Almost all the graphs show a gradual rise with the increase of the number of input syllables and no significant difference between the five formats of f_0 input except the input of f_0 velocity which always had slightly worse performance than other formats of f_0 input. For the whole sentence length (5 syllables), F_Syl1 always has better performance than F_Syl2, and models that consider f_0 context for local focus events have higher accuracies (>80 %) than those without considering f_0 context (>70 %) with F_Syl1.

3.2.2. Accuracy of individual focus categories

Fig. 12 (a–e) shows accuracies of the four focus categories with f_0 in Hz as input with different sentence fragments tested. The left and right columns display results based on F_Syl1 and F_Syl2 labeling schemes, respectively. The two labeling schemes differ extensively, but always show similar patterns for Initial focus. At the whole sentence length, F_Syl1 always has a similar accuracy ranking: Middle focus > Initial focus > Neutral focus > Final focus. The ranking for F_Syl2, however, is almost always: Initial focus > Neutral focus > Middle focus > Final focus.

Referring to the perceptual results (Fig. 22 in Section 3.6), F-GRU and F-Bi-GRU had the most similar patterns when using F_Syl1 labels. Looking at Fig. 12(b), from two-syllable fragment to whole sentence, the accuracy of neutral focus decreased from 97.4 % to 85.94 % to 71.35 %, and the accuracy of initial focus increased from 75 % to 79.65 % to 89.06 %. In the two-syllable condition, the rate of middle focus is 0 %. After that, when focused word was added, the rate of middle focus suddenly increased to 91.15 % and then climbed further to 96.35 % at whole sentence condition. Final focus always had the lowest accuracy, staying at 0 % when the final word was missing, and reaching above chance only at whole sentence condition.

3.3. Experiment 3: recognition of tone-focus combinations

The recognition of tone-focus combination is a way to simulate the consideration of the effect of co-occurring function, which is presumably beneficial. In general, however, the performance of the tone-focus combination models achieved similar results as models that recognize tone and focus independently of each other in Experiments 1 and 2.

3.3.1. Tone accuracy

Fig. 13 shows UARs of tones based on the five tone-focus combination models with five formats of f_0 input. The overall tone accuracies of the first three models are all above 85 % and the accuracies of the last two models both reached over 95 %.

Fig. 14 shows the UARs of tone based on TF-GRU-Con and TF-Bi-GRU-Con tested with sentence fragments from two to five syllables. Both of the models only had chance-level accuracy when tested with only the second syllable. With f_0 of the first two syllables, the performances increased dramatically. TF-GRU-Con achieved the same performance under the rest of the three conditions with accuracies around 95 % (f_0 in Hz). TF-Bi-GRU-Con had slight increase with more input syllables, with accuracy rising from 90 % to 97 % (f_0 in Hz).

Fig. 15 shows accuracies of the four tones with f_0 in Hz as input based on the five combination models. Same as tone only recognition, all the five models showed that static tones had higher accuracies than dynamic tones: T1 > T3 > T4 > T2. The first three models show lower performance than the last two models.

3.3.2. Accuracy of individual focus categories

Fig. 16 (a–e) shows the overall accuracies of focus recognition from

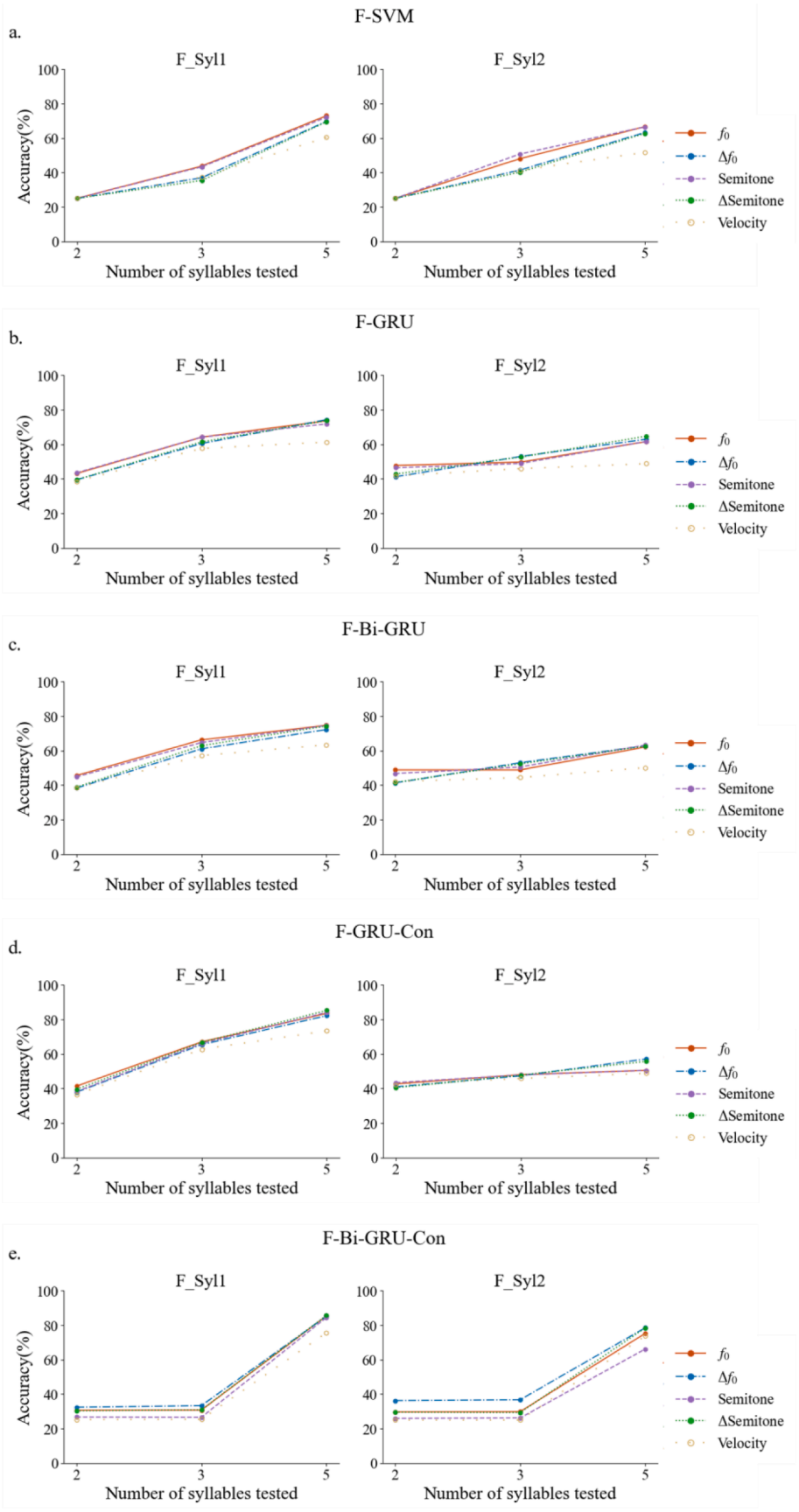

Fig. 11. Overall accuracies of sentential focus as a function of the number of tested syllables from the five recognition models with F_Syl1 (left) and F_Syl2 (right).

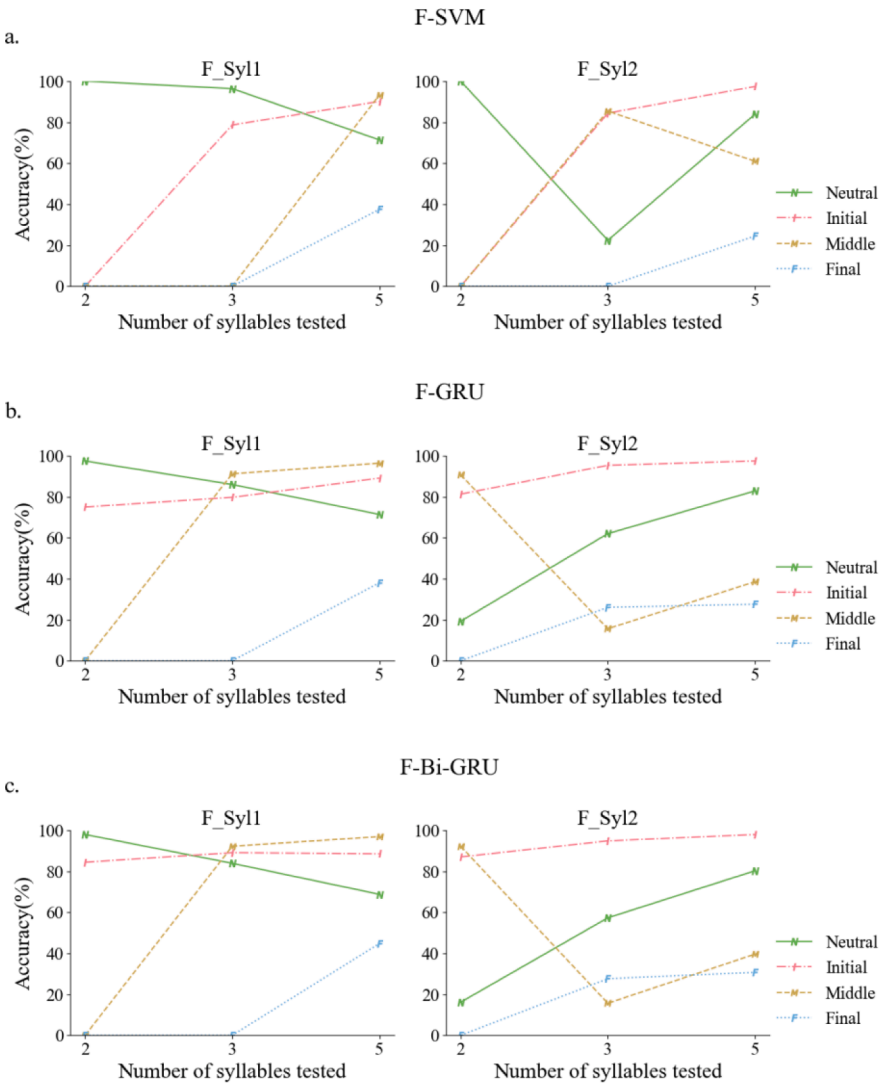

Fig. 12. Recognition rates of the four focus categories as a function of the number of tested syllables from the five recognition models with F_Syl1 (left) and F_Syl2 (right).

the five tone-focus combination models. The left column displays results based on F_Syl1 labels and the right column displays results based on F_Syl2 labels. Similar to syllable-by-syllable focus only recognition, almost all the graphs show a gradual rise with the increase in the number of input syllables and no significant difference was found between the different formats of f_0 input except the input of f_0 velocity which always had slightly worse performance than other formats of f_0 input. Unlike in Experiment 2, the models trained with F_Syl1 labels did not outperform those trained with F_Syl2 labels. But similarly, models considering f_0 context for syllabic events had higher accuracies (>80 %) than those not considering context information (>70 %) at whole sentence condition.

Fig. 17 (a–e) shows the accuracies of four sentential focus with f_0 in Hz as input. The left column displays results based on F_Syl1 labels and the right column displays results based on F_Syl2 labels. The models using F_Syl1 labels have similar results as the syllable-by-syllable focus only recognition, and TF-GRU and TF-Bi-GRU had the most similar patterns to perceptual results as shown in Fig. 22 when they were trained with F_Syl1 labels. The performance of models with F_Syl2 showed some differences from the syllable-by-syllable focus only tasks in Experiment 2, but the initial focus sentences still have the same pattern when trained with the two label sets.

3.4. Experiment 4: holistic focus recognition

3.4.1. Overall accuracy

Fig. 18 shows the overall focus recognition accuracies of the three holistic models. All the models show a small increase from the first 2 syllables to the first 3 syllables, and then a large increase from 3 syllables to 5 syllables. There is no significant difference between different formats of inputs f_0 data except f_0 velocity.

3.4.2. Accuracy of each sentence focus category

Fig. 19 shows the accuracies of each focus categories by SF-SVM (left), SF-GRU (middle) and SF-Bi-GRU (right) with f_0 in Hz as input.

3.5. Tone perception by listeners

The perceptual data were obtained from 19 of the 20 listeners who participated in the experiment. The excluded subject was a male who performed abnormally on tone 3 (nearly 0 %). Fig. 20 shows the accuracies of identifying the four Mandarin tones on the second syllable across the four fragmented sentences: the second syllable only, the first word, the first two words and the whole sentence. The overall accuracies are 78.69 %, 86.24 %, 87.68 % and 88.15 %, respectively.

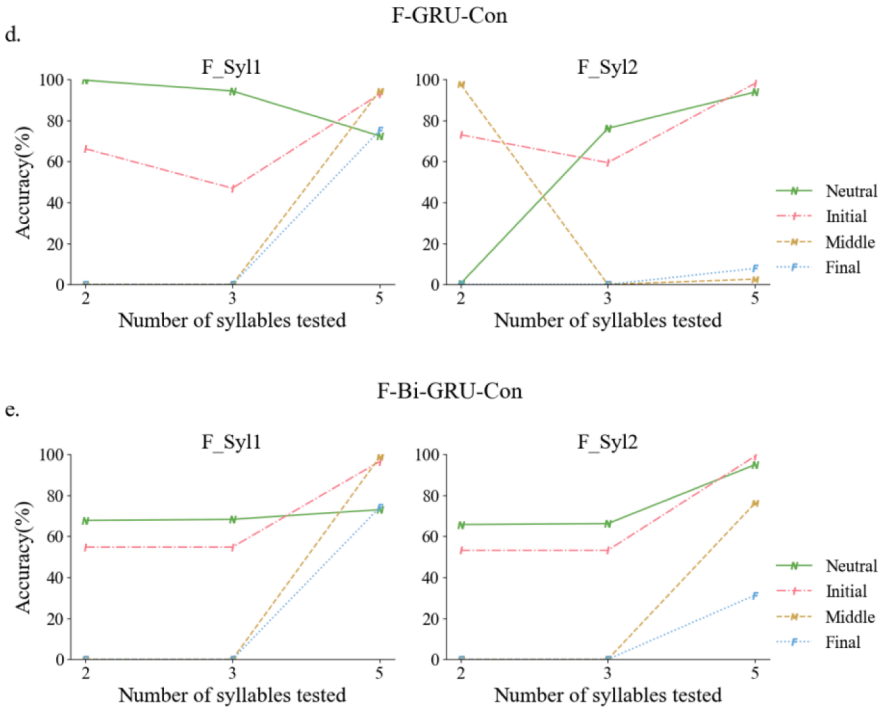

Fig. 12. (*continued*).

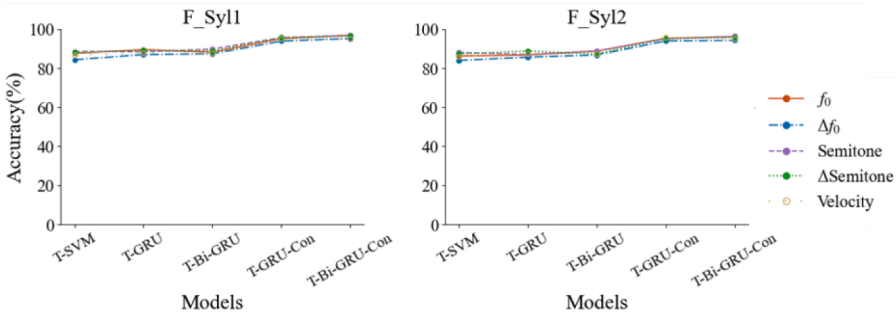

Fig. 13. Overall UARs of tones based on the five tone-focus combination recognition models.

As shown in Fig. 20, in most cases, the accuracies are ranked as T1 > T3 > T2 > T4. The static tones are always better perceived than dynamic tones. The accuracies of T1 are always high, while those of T2 and T3 increase from one syllable condition to one word condition and to two words condition. T4 always had the lowest accuracy around 65 % except when listeners heard only the first word where the accuracy reached 80 %.

Fig. 21 shows heatmaps of confusion matrices for tone identification when listeners heard the whole sentence, the first two words, the first word, and the second syllable only, respectively.

It is noteworthy that, unlike the popular claim that T2 and T3 are easily confused with each other (Lee et al., 2008; Shen and Lin, 1991), here T2 is more likely confused with T1, and T4 is more likely confused with Tone 3.

3.6. Focus perception

The overall accuracies of focus identification are 28.84 %, 29.04 %, 55.53 % and 67.81 % under the conditions when listeners heard the second syllable, the first word, the first two words, and the whole sentence, respectively. Fig. 22 shows the perceptual accuracies of the four focus categories with different f_0 fragment sizes. It can be seen that except for neutral focus, all other focus categories increased their accuracies as the fragmented f_0 contour became longer, and the accuracies were 0 % when focused words were not included. Final focus had the lowest accuracies among all the conditions.

Fig. 23 shows the heatmaps of confusion matrices for focus identification when listeners heard f_0 contours with different fragment sizes.

In all the four confusion matrices, neutral focus is always more likely to be misperceived as initial focus, whereas all the other focus categories are likely to be misperceived as neutral focus.

4. Discussion

The present study is aimed at using computational simulation to explore how multiple melodic functions in speech, e.g., tone and focus, can be recognized both simultaneously and sequentially in perception. For this purpose, we tried to answer two specific research questions: 1) Can tone and focus be recognized independently of each other, or they have to be co-processed for concurrent recognition? 2) Does focus have to be recognized by processing sentence-wide f_0 contour as a whole, or the recognition can be done syllable-by-syllable, guided by progressive accumulation of probability? To answer these questions, we trained several tone and focus recognition models, each designed to simulate one or more hypothetical strategies. In general, the results of the modeling show evidence that a) tone and focus can be processed largely

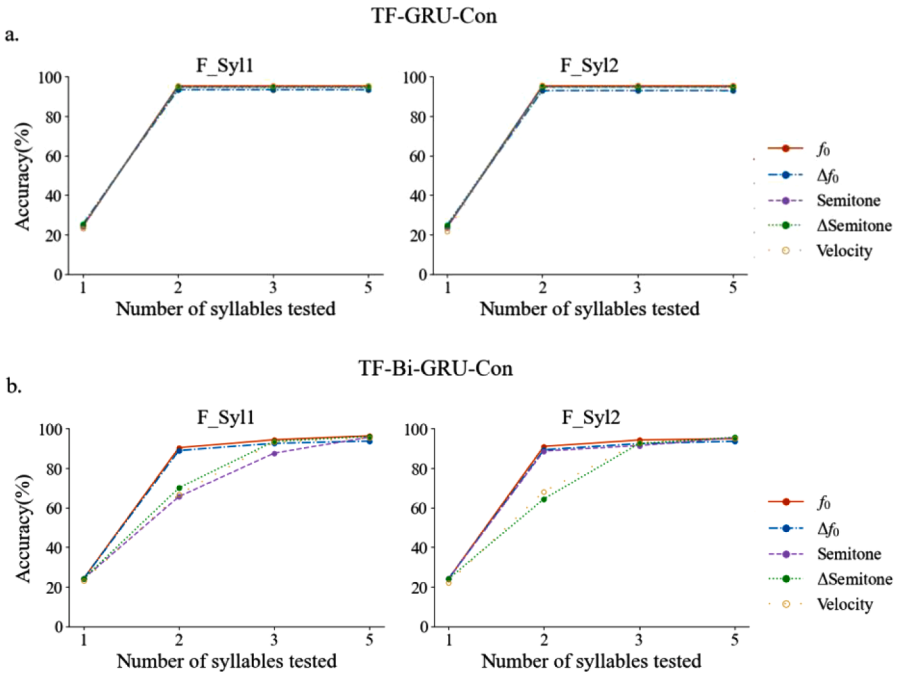

Fig. 14. Overall UAR of tone under combination recognition as a function of the number of tested syllables from TF-GRU-Sen and TF-Bi-GRU-Syl.

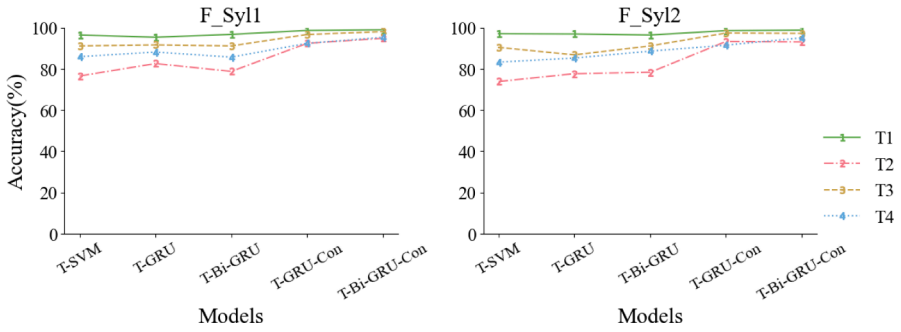

Fig. 15. Recognition rates of the four tones based on the five combination recognition models tested with syllables of whole utterances.

independently for their respective recognition, and b) focus can be recognized syllable-by-syllable from continuous speech. Before discussing the main results in detail, however, we would like to first remark on a finding that is almost a minor byproduct of the study, namely, the nonessential role of explicit f_0 normalization in the perceptual decoding of tone and focus.

4.1. *No benefit of f_0 normalization*

In all the simulations, raw f_0 trajectories, except for time-normalization, were used as training and testing data, without the extraction of intermediated features, following the finding of Chen et al. (2022). They all achieved recognition accuracies no worse than human perceptual performance on the same data, despite the fact that the human listener could have benefited from non-f_0 cues in the stimuli that are unavailable to the models. In addition, the transformations from the original f_0 in Hz to other formats (semitones, Δf_0, and velocity) to remove most of the overall pitch height differences across speakers and utterances did not show any benefit. This unexpected finding suggests that, raw f_0, despite being full of variability that is widely assumed to be in need of perceptual normalization (Francis et al., 2006; Johnson and Sjerps, 2021; Wong and Diehl, 2003), in fact carries plenty of useful information for the perceptual identification of both tone and focus. It is likely the case that, thanks to the very need for sufficient intelligibility, the encoding schemes of important communicative functions, including tone and focus, may have been shaped in such a way in daily conversations that their distinctive cues would rise above or cannot be masked by the most commonly occurring variability. But this possibility is hard to demonstrate by behavioral studies alone. What the current modeling results have demonstrated, following our previous finding in Chen et al. (2022), is that speaker variability probably is never a formidable barrier as is often assumed, and normalization, even if necessary, can be easily achieved as part of the training/learning process itself rather than as an extra pre-processing. Theoretical implications notwithstanding, in the following we will only discuss the results based on f_0 in Hz given its equivalent and often superior performance to the other f_0 formats.

4.2. *Syllable-by-syllable processing of speech melody*

4.2.1. *Processing prosodic focus syllable-by-syllable*

The results of human perception of focus show a general improvement as more and more fragments of the sentence were included in the stimuli, which support the idea that intonation can be perceived progressively. The overall accuracies of the recognition models trained by both syllable-by-syllable and holistic f_0 contours also showed gradual upward trends with increased sentence fragment size. As shown in Figs. 22 and 23 (in Section 3.6), at the beginning of the sentences, listeners tend to assume that the sentence has neutral focus. Once the

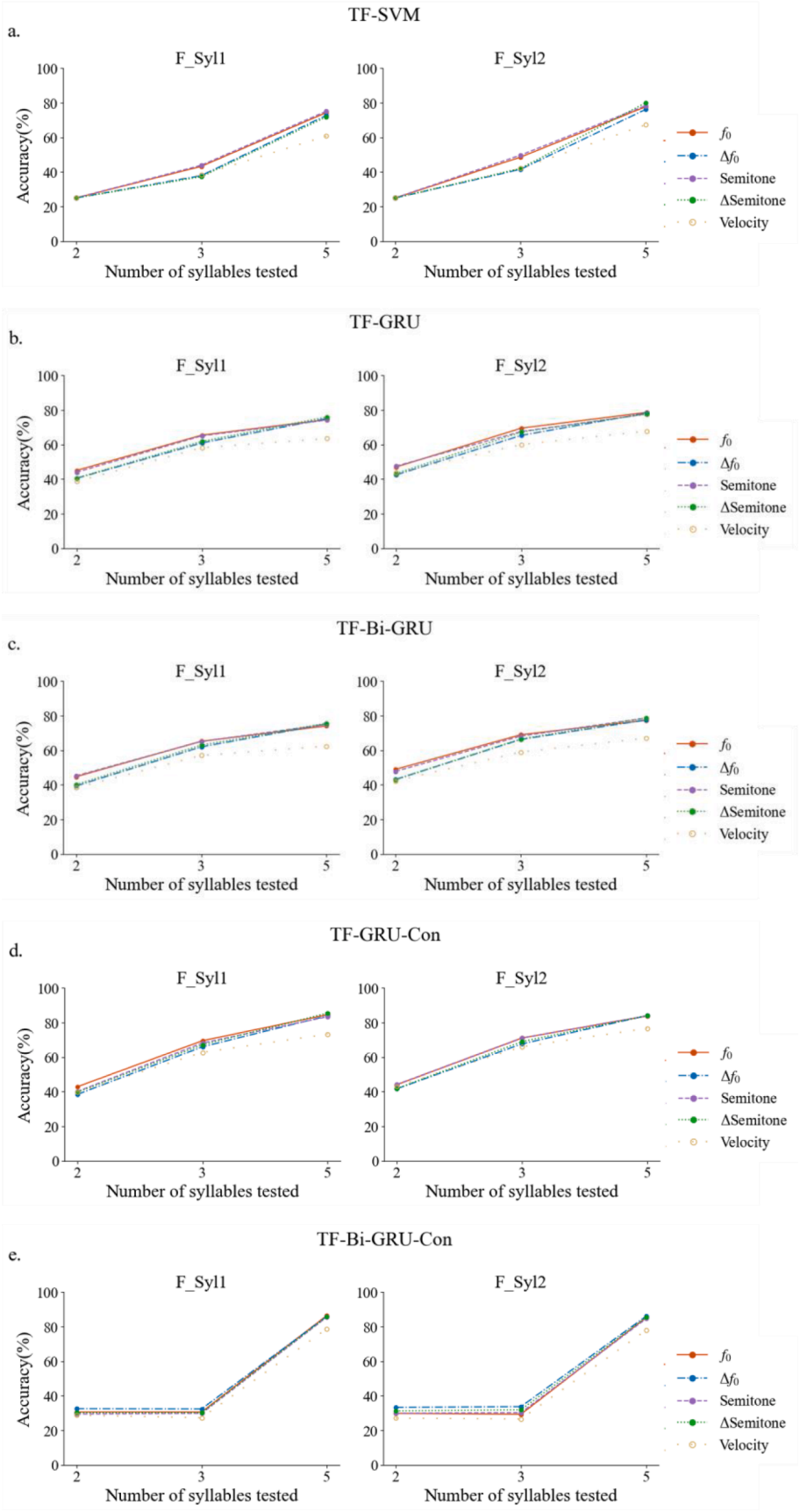

Fig. 16. Overall accuracies of sentential focus as a function of the number of tested syllables from the five combination tasks with F_Syl1 (left) and F_Syl2 (right).

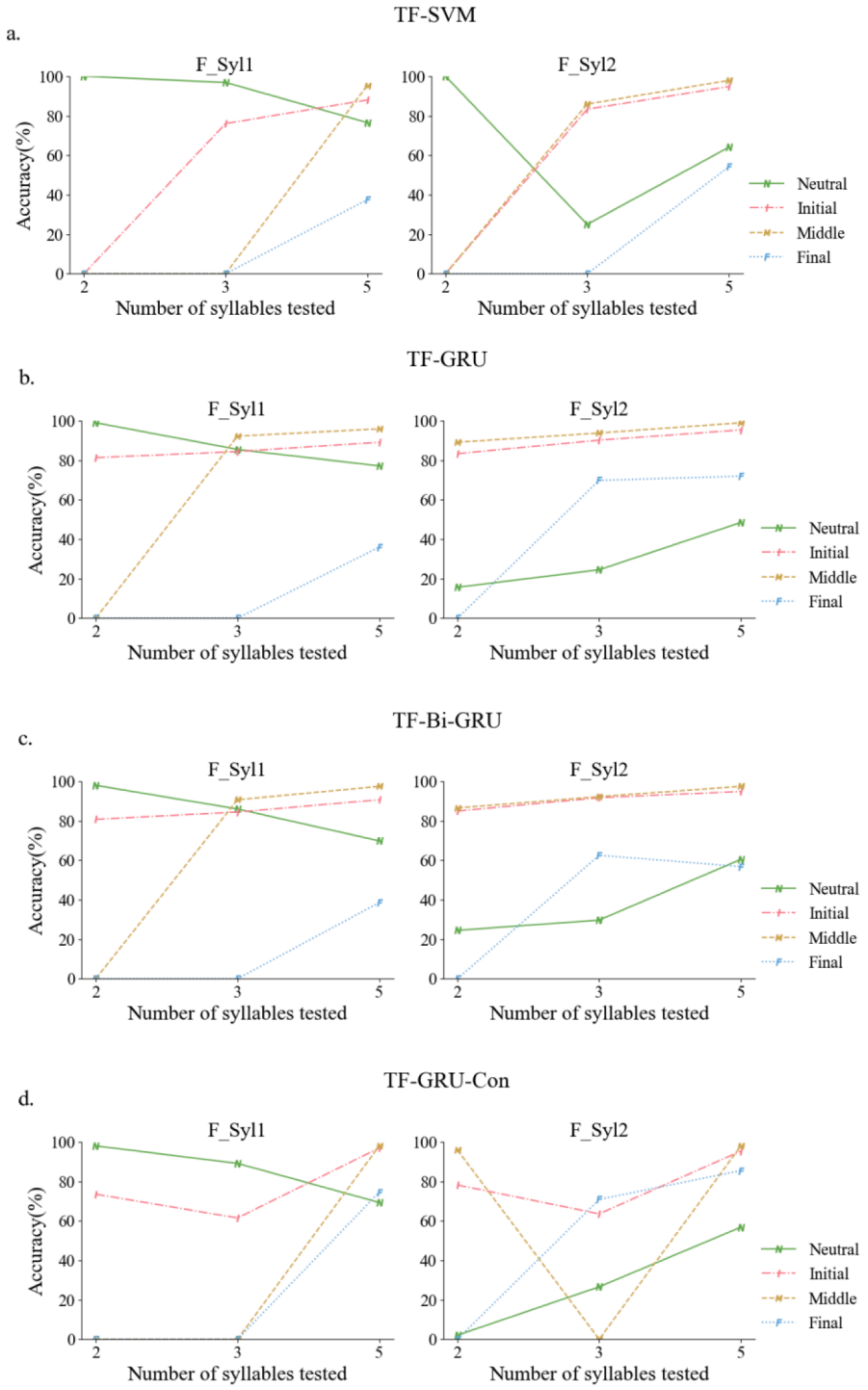

Fig. 17. Recognition rates of the four focus types as a function of the number of tested syllables from the five combination tasks with F_Syl1 (left) and F_Syl2 (right).

focused word was heard, the perceptual accuracy of focus increased rapidly, and the ambiguity was further reduced when post-focus words were heard. In contrast, the accuracy of neutral-focus sentences dropped gradually as more syllables were heard. For the whole sentences, medial focus was perceived the best, and final focus was perceived the worst. This is consistent with focus perception results from Yuan (2011) and Liu (2009), where in statement, the overall accuracies of focus at different positions are ranked as: Middle focus > Initial Focus > Final focus.

For the computational simulations, the holistic models did not generate recognition patterns similar to human perception, especially showing no bias toward neutral focus (Fig. 19 in Section 3.4.2). In contrast, syllable-by-syllable models did show focus recognition patterns similar to human perception, especially when trained with F_Syl1 labels (Fig. 12(b) in Section 3.2.2, F-GRU with label set F_Syl1). Looking at the classification performance on syllable-level focus events (Fig. 24), the syllable-by-syllable models showed significant distinctions between syllables in different portion of the sentence-wide tri-zone profile. Post-focus syllables had the highest accuracy above 80 %, which seems to be the most robust cue for focus recognition than the focused syllables,

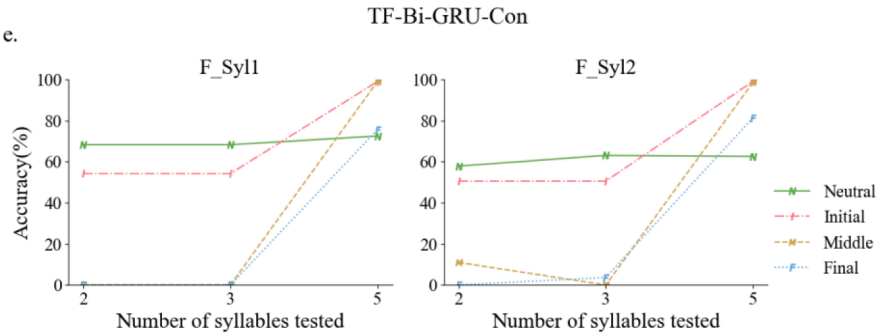

Fig. 17. (*continued*).

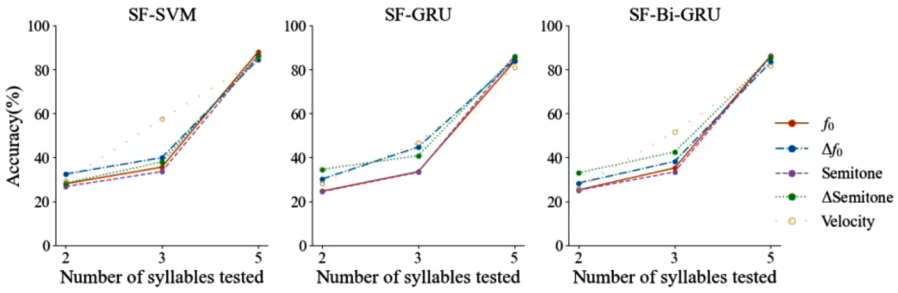

Fig. 18. Overall focus recognition accuracy as a function of the number of tested syllables from SF-SVM (left), SF-GRU (middle) and SF-Bi-GRU (right).

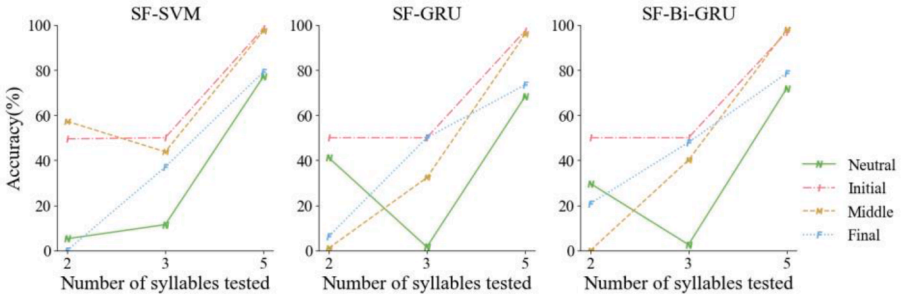

Fig. 19. Recognition accuracies of the four focus types as a function of the number of tested syllables from SF-SVM (left), SF-GRU (middle) and SF-Bi-GRU (right).

regardless of the tone composition of the sentence. This is consistent with previous focus perception experiments (Ipek, 2011; Rump and Collier, 1996; Xu et al., 2012). The on-focus syllables had the lowest accuracy, although still above 50 %. When on-focus and final on-focus syllables were treated as two different categories, the on-focus accuracy increased to over 70 %. This could be because the declination effect on the final word neutralizes the on-focus pitch expansion which makes the final on-focus syllables behave like neutral focus syllables.

A noticeable difference between perception and recognition (F-GRU) patterns is that with sentence fragment of the first word (two syllables), the perceptual rate of initial focus is only around 30 % while the recognition rate is much higher at 75 %. When the following word was given to listeners, the accuracy of initial focus rose rapidly to 70 %. Listeners' insensitivity to focused syllables at initial position is possibly because listeners are not familiar with the speakers. Also, the first syllable in the sentences is always high tone (T1) which leaves limited space for focus encoding. Thus, a contextual environment is probably needed to show the pitch prominence. The computational models, comparatively, are much more familiar with the corpus, and could identify the initial focused words immediately and accurately.

4.2.2. *Context-free local recognition of tone and focus from individual syllables*

For the tone recognition models, we tested two f_0 processing strategies: processing only syllable-sized f_0 contours, or also taking surrounding (T-Bi-GRU-Con) or just preceding (T-GRU-Con) f_0 contours into consideration. The goal was to assess the benefit of context dependent processing of f_0 given previous findings of an important role in tone perception (Gottfried and Suiter, 1997; Lee et al., 2008; Xu, 1994). The effect of context was also assessed in the perception experiment by comparing f_0 contours with different fragment sizes. The tone perception results showed that even without context, the identification of the tone of the second syllable was fairly accurate, with only moderate drops for T2 and T3 (Fig. 20 in Section 3.5). In the computational simulation, the context dependent models (T-GRU-Con & T-Bi-GRU--Con) performed badly when tested with all contextual f_0 replaced with a grand average (Fig. 9 in Section 3.1.1). The context-free models, in contrast, could recognize tones not only very well (89 %), but also better than human perception of the second syllable without context (79 %). The reason could be that f_0 contours of individual syllables do contain sufficient cues for distinguishing the tones in this corpus. The human subjects in the study, however, have rarely been exposed to tones extracted from context which are as short as those of the second syllable in the current corpus. If this interpretation is valid, it can be viewed as

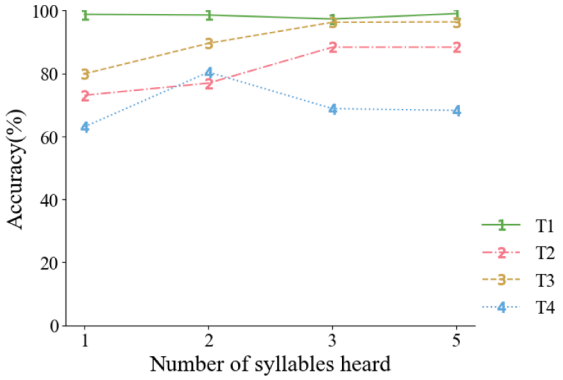

Fig. 20. Perceptual accuracy of the four tones as a function of number of syllables included in the stimuli.

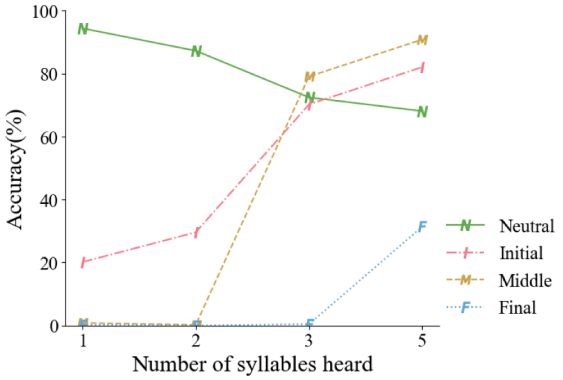

Fig. 22. Perceptual accuracy of the four focus categories as a function of number of syllables included in the stimuli.

evidence of the role of perceptual learning, which is also how speaker learn to process tones produced by different speakers, which they would have had plenty of exposures in their listening experience. Yet another possible reason for the poorer human perception of the tone in isolation compared to the context-free models is that our current syllable segmentation is not correct, given the latest finding of the much earlier alignment of tone as well as both the consonant and vowel in a syllable (Kang and Xu, 2024; Liu et al., 2022).

For focus recognition, the context-dependent models (F-GRU-Con & F-Bi-GRU-Con) achieved higher accuracies of syllabic events when tested on whole sentences, but they did not show the same confusion and incremental patterns as human focus perception. In contrast, the context-free models (F-GRU & F-Bi-GRU) trained with F_Syl1 labels showed focus recognition patterns (Fig. 12(b), left column Section 3.2.2) much more similar to human perception (Fig. 22 in Section 3.6) as the sentences unfolded over time.

The syllable-by-syllable accumulation of evidence in the simulations was done through Bayesian inference and the local recognition probabilities were integrated so as to identify the focus of the sentence. This approach can utilize all the cues in the utterance in a more flexible and efficient way than holistic recognition. Also, as a context-free model, F-SVM achieved similar performance to the models using GRU at syllable level, but did not work well at sentence level, with a late response to local focus events (Fig. 12(a) in Section 3.2.1). It is because SVM could not give confidence intervals directly and the probabilities are calibrated using logistic regression on the SVM's scores (Platt, 2000; Wu et al., 2004). Thus, the probabilities obtained from the SVM are not as reliable as neural networks like GRU. This further highlights the benefit of GRU and the integration of gradient cues that occur at different points in time. That is, the more accurate the local perceptual probabilities, the better the final decision regardless the context information.

The success of local recognition of tone and focus and the later syllable-based inference for sentence focus thus demonstrates that syllable-sized unit could be used as the smallest temporal scope of speech melody processing at least in Mandarin.

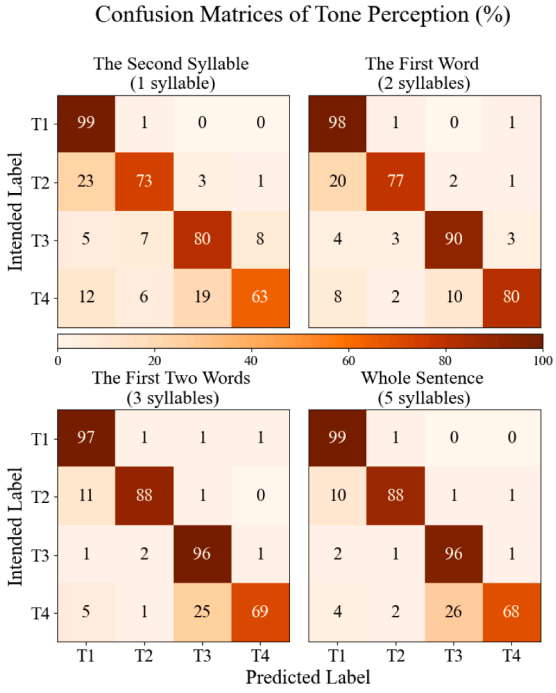

Fig. 21. Heatmaps of confusion matrices for perceptual identification of the tone of the second syllable when listeners heard the second syllable (mi) only, the first word (mao mi), the first two words (mao mi mo), and the whole sentence (mao mi mo mao mi), respectively.

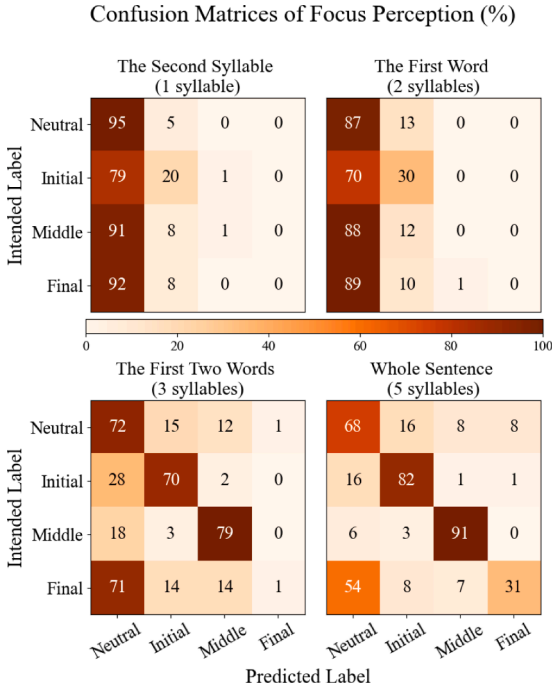

Fig. 23. Heatmap confusion matrices for perceptual identification of focus when listeners heard the second syllable (mi) only, the first word (mao mi), the first two words (mao mi mo), and the whole sentence (mao mi mo mao mi), respectively.

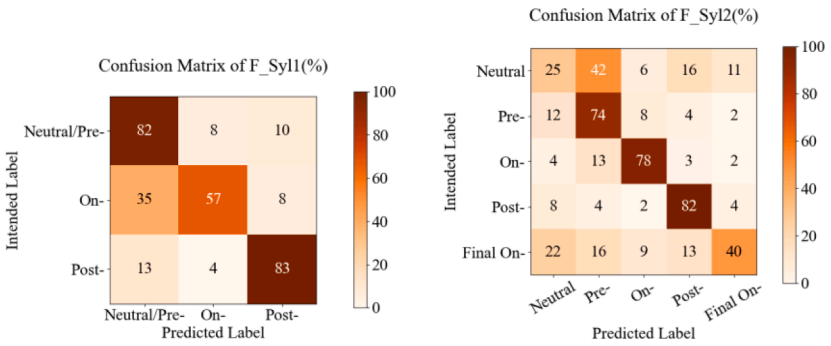

Fig. 24. Heatmap confusion matrices for syllable-level focus events recognition by F-GRU model trained with F_Syl1 and F_Syl2 labels, respectively.

4.2.3. *Default bias toward neutral focus*

The two focus labeling schemes in the present study were used to examine two unresolved issues. The first is the claim that there is a default focus location at the end of each sentence in Mandarin (Yan and Calhoun, 2020) as well as in English (Büring, 2006; Carlson et al., 2009; Ladd, 2008), which may imply that even in the absence of a narrow focus listeners would still hear a sentence-final focus. The second is the uncertainty as to whether there are prosodic cues to mark pre-focus words as distinct from neutral focus words. Empirical findings on this issue are mixed (yes: Alzaidi et al., 2019; no: Xu, 1999; Xu and Xu, 2005).

To address the first issue, F_Syl2 assigns different labels for sentence-final focus and non-final focus: Final On- vs. On-. This was to maximize the recognition rate of final focus by allowing it to be trained separately from non-final focus which has higher overall f_0 than final focus (Xu, 1999). This is based on the fact that final focus is already known to be less salient than non-final focus (Botinis et al., 1999; Xu et al., 2012). As shown in Fig. 24, many final on-focus syllables were recognized as neutral focus (21.88 %), whereas only a few neutral focus syllables were recognized as Final On- (10.52 %). This is also consistent with Fig. 12 (in Section 3.2.2) where final focus always had the lowest recognition accuracy. Also, human focus perception showed no final-focus bias neither (Fig. 23 in Section 3.6). For the whole sentence stimuli, neutral focus was misidentified more as initial focus rather than as final focus, whereas final focus was mostly heard as neutral focus (54.25 %) rather than as itself (31.38 %). These results are the opposite of the notion of broad focus (Ladd, 2008) or default sentence-final focus (Büring, 2006; Carlson et al., 2009; Yan and Calhoun, 2020). Instead, at least for Mandarin, the default focus status of a sentence is neutral rather than final or broad.

To address the second issue, F_Syl2 labeled pre-focus syllables as different from syllables in a neutral focus sentence: Pre- vs. Neutral. The results of the F-GRU model in Fig. 12 show that the major differences between F_Syl1 and F_Syl2 are in the accuracies of neutral focus and middle focus in the two-syllable and three-syllable conditions. With F_Syl1, neutral focus accuracy is high at 97 % at two syllables, but drops to 86 % at three syllables, while the accuracy of middle focus increases from 0 % to 91 %. With F_Syl2, the trends are reversed. As shown in Fig. 24, the misidentified neutral focus syllable is more biased toward pre-focus syllable. As pre-focus syllables are always in the first half of the sentences and neutral focus syllables are distributed across the whole sentences, given the general f_0 declination of the sentence (Xu, 1999), pre-focus would have higher overall f_0 in general and neutral focus would have more variations. This would lead to a bias in the trained model to identify neutral-focus syllables with higher f_0 as pre-focus. But such difference between neutral-focus and pre-focus syllable is not seen from human listeners (Fig. 23 in Section 3.6). They did not predict any upcoming focus when hearing only pre-focus words. Combining the results of model simulation and human perception, therefore, pre-focus words do not seem to carry much essential focus information.

Taken the findings about both issues together, there seems to be a clear bias toward neutral focus, such that effective focus encoding does not start until the focused word, which is further enhanced by post-focus compression. When there no strong reason to emphasize any particular component of a sentence, the utterance is simply spoken without any focus. When the final word of sentence needs to be emphasized, due to the lack of subsequent words, the focus cue is compromised, resulting in weakly encoded final focus.

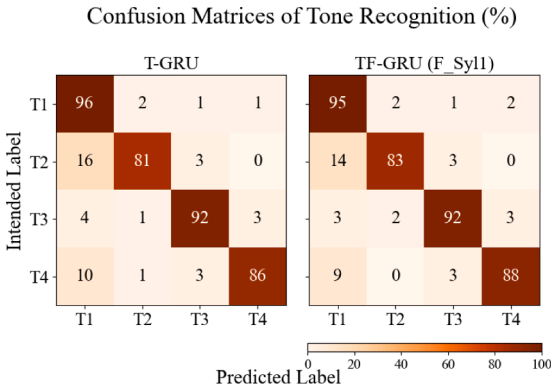

Fig. 25. Heatmap confusion matrices of tone recognition models of T-GRU and TF-GRU tested on syllables from whole sentences.

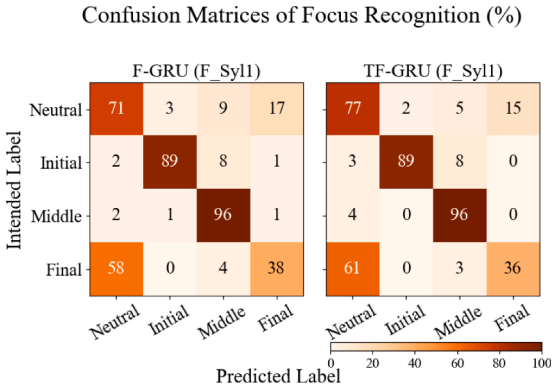

Fig. 26. Heatmap confusion matrices of focus recognition models of F-GRU and TF-GRU tested on whole sentences.

4.3. *Independent recognition of tone and focus*

The second research question was mainly addressed by Experiment 3, which explicitly simulated the co-decoding/co-processing of tone and focus by recognizing categorical tone-focus combinations from syllable-sized f_0 contours. Somewhat surprisingly, however, no clear difference in recognition accuracy was found between the tone-focus combination

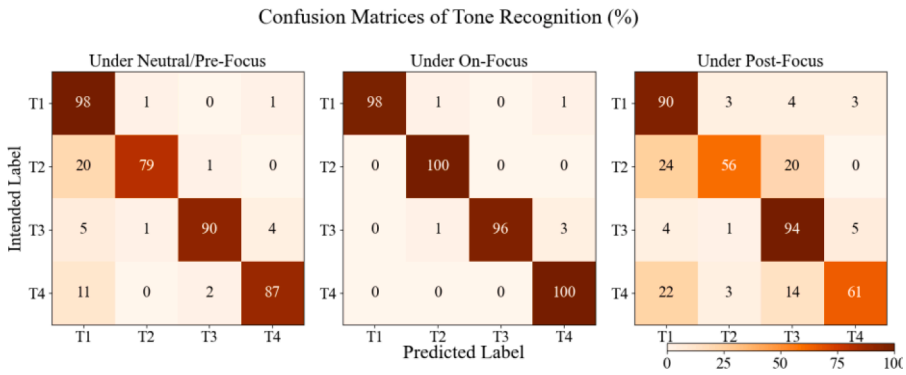

Fig. 27. Heatmap confusion matrices based on breakdown analyses of tone recognition across different local focus events of T-GRU.

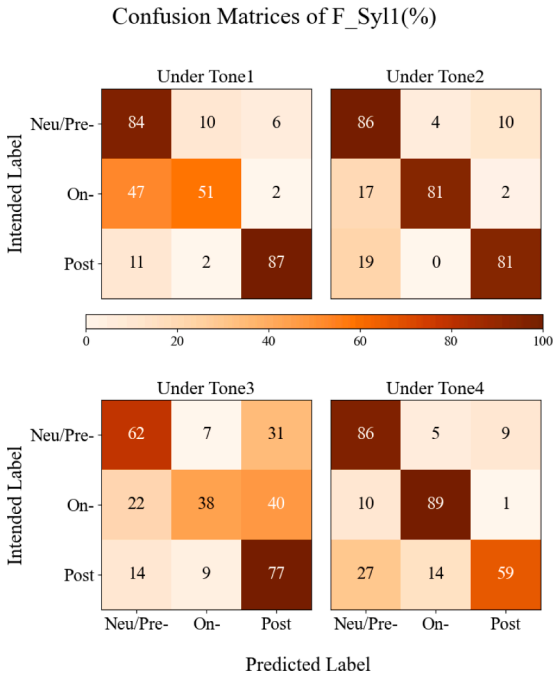

Fig. 28. Heatmap confusion matrices base on breakdown analyses of focus recognition across different tone conditions of F-GRU with F_Syl1.

task and the independent recognition tasks.

4.3.1. *Similarity between single function recognition and combination recognition*

Figs. 25 and 26 compare the confusion matrixes of single function recognition (T-GRU and F-GRU) and combination recognition (TF-GRU) with F_Syl1 labeling scheme. The results of single function and combination recognition are very similar to each other and the confusion patterns are also largely in accord with perceptual results. For tone, T1 always has the highest accuracy in all confusion matrices in Fig. 25, followed by T3. Accuracies of T2 and T4 are much lower. T2 is more easily confused with T1 rather than with T3. Interestingly, in Fig. 21, T4 has a much lower perceptual accuracy than the other tones when listeners were given the whole sentences. But this differs from both of the recognition models. From Fig. 21, it can be seen that most of the perceptual confusion of T4 was with T3. It seems that listeners were attending to the lowest f_0 reached *after* the T4 syllable, which was due to the carryover effect of inertia (Xu, 1997, 2005). If this is the case, it also means that they did not take the syllable boundary very seriously, or they could not perform syllable segmentation with high accuracy. This interpretation is supported by the fact that when they heard only the first two syllables, which did not include the lowered f_0 due to the carryover effect, the accuracy of T4 rose up to 80.4 %. The perceptual confusion pattern becomes more similar to the performance of tone recognition tasks in Fig. 25.

As for focus identification, the overall perceptual accuracy (Fig. 23) is lower than model recognition accuracy (Fig. 26), and most of the perceptual inaccuracies are due to confusion with neutral focus. One possible reason is that human listeners have never been trained to recognize focus as intensively as our models. But the key similarity between focus-only recognition and tone-focus combination recognition nevertheless is the same as in Fig. 25, showing no advantage of processing f_0 contours explicitly as tone-focus combinations.

The finding of lack of advantage of tone-focus combination is surprising because tone and focus are indeed fused with each other in the f_0 contours, and we have been able to simulate their parallel encoding in production through syllable-sized singular target that encode both tone and focus using PENTAtrainer (Xu and Prom-On, 2014). What the new finding suggests is that tone and focus likely use separate encoding spaces and/or dimensions despite doing so both through f_0, and perception training can learn to decode them separately.

4.3.2. *Interactions between tone and focus*

The lack of overall difference between separate tone and focus recognition and the recognition of tone and focus as combinations does not mean that the two functions do not affect each other. Their mutual influence on each other can be seen in Figs. 27 and 28. Fig. 27 shows clear tri-zone focus effects on lexical tones: highest accuracy under focus, median accuracy before focus and in neutral focus, and lowest accuracy after focus, especially for dynamic tones (T2 and T4).

Fig. 28 shows that on-focus has lower recognition accuracy under T1 and T3 than under T2 and T4. T1 and T3 are both static tones which may have left limited space for distinctive on-focus pitch range expansion. Especially when T3 is in the final focused word, pitch expansion result in even lower pitch, which makes it confusable with post-focus compression, especially for the final-focused words whose first syllable is in T3.

Importantly, all these interactions occurred both with T-GRU and F-GRU and with TF-GRU, which means that the interactions (the combination of tone and focus labels) affected the degrees of ambiguity of both tone and focus, but it is not helpful to recognize tone and focus together. In other words, therefore, there is little need to take focus into consideration when recognizing tone, and vice versa.

Also, from the breakdown analyses in Figs. 27 and 28, the effects of tone on focus are greater than the other way around. The overall accuracies of tones are higher than focus and the tone recognition patterns under each focus location are fairly stable. Such more robust f_0 encoding of tone than focus is a bit surprising given that focus uses a larger pitch range than tone. But it is also not that surprising given the early finding of (Fry, 1958) that only a 5 Hz f_0 difference is sufficient to elicit categorical perception of lexical stress in English, and that larger f_0 differences did not lead to any further improvement in stress perception. It

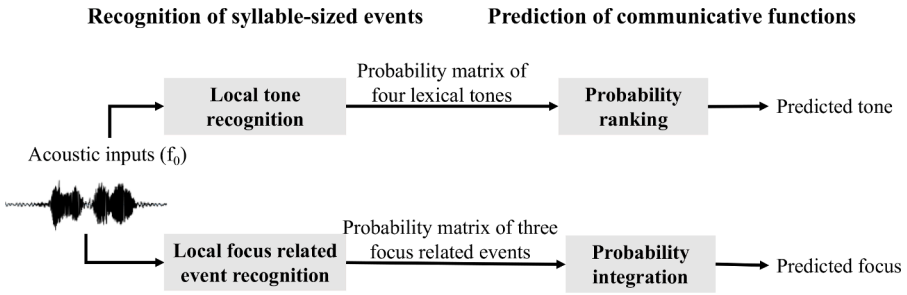

Fig. 29. A framework of tone and focus perception from continuous speech.

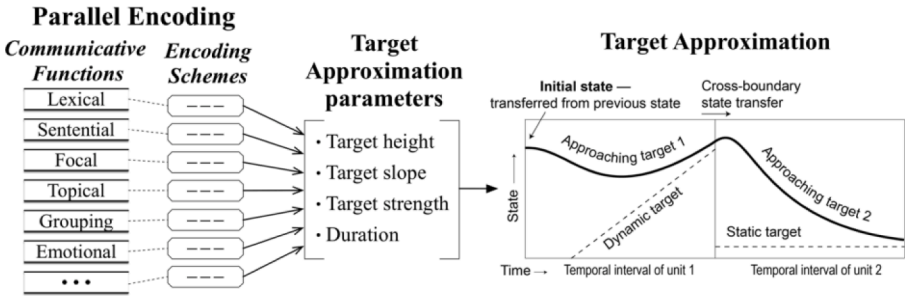

Fig. 30. The PENTA model of speech melody production (Xu, 2005). The communicative functions (block 1 from left) are encoded in parallel through encoding schemes (block 2) that are specified in terms of target parameters (block 3), which in production are articulated via target approximation to generate continuous f_0 contours (block 4).

could be the case that the greater accuracy of tone encoding is because lexical contrasts occur in every word while focal contrasts occur much less frequently, as indicated by listeners' bias toward hearing neutral focus as discussed above. But a clear understanding of this surprise finding has to wait for further studies.

4.4. Summary and conclusion

In this study, we used computational modeling to explore whether it is possible to simulate simultaneous recognition of tone and focus in Mandarin by processing f_0 contours syllable-by-syllable. We tested this possibility by training SVM and GRU models with either syllable-sized or sentence-sized f_0 contours from an existing corpus, and testing them with f_0 contour fragments at different sizes and locations in the sentence. The recognition accuracies of these models were then compared to human perception accuracies on stimuli from the same corpus. The main results are as follows.

1. The most comparable simulations were seen from GRU model trained with syllable-sized f_0 contours, and recognized both tone and focus by progressive accumulation of local probability.
2. Models trained with sentence-sized global f_0 contours did not recognize focus as well as those trained with syllable-sized f_0 contours.
3. There was little difference whether tone and focus were recognized separately or as tone-focus combinations in model simulation, suggesting that tone and focus can be recognized independently of each other.
4. There was evidence of a clear preference for neutral focus as the default focal category in human perception as well as our model simulation.
5. No clear benefits were found for taking f_0 context into consideration even for tone recognition.
6. No advantages were found for applying various f_0 normalization schemes that may filter out speaker differences and random variations.

Of these findings, only the first was the main hypothesis that we had hoped to corroborate at the outset of the study. For the others, we either did not have a strong prior preference (2–4) or even expected the opposite (5–6). Based on the first 3 findings, we can offer a schematic framework of tone and focus perception shown in Fig. 29. The framework is based on the assumption that speech melody conveys multiple layers of communicative functions which are articulatorily encoded in parallel through syllable-sized f_0 contours (Xu, 2005), as illustrated in Fig. 30. However, the decoding process in Fig. 29 is not a direct reverse of the encoding process in Fig. 30. Rather, it is a simpler process of identifying the component elements of each function in separate parsing routes. On the other hand, the decoding does proceed syllable by syllable as they are the most relevant chunks of encoding events generated by production (Xu, 2020; Xu and Prom-On, 2014). For functions like focus whose temporal scopes transcend the syllable, their identifications are done in steps by progressively updating the probability at each syllable before reaching a final recognition at the end of the function's temporal scope.

It is truly surprising that the consideration of f_0 context did not generate much benefit when processing syllable-sized f_0 contours not only for focus but also for tones whose encoding is largely local. It could be the case that the most informative property of each f_0 contour is its movement toward the underlying target (Xu, 2005). But not taking the consequence of the target approximation on the context is surprising. This, however, will need further confirmation in future studies.

In summary, what the present study has demonstrated is that the perception of tone and focus in Mandarin can be done in parallel, but not through reversing the articulatory encoding by identifying tone-focus combinations generated by joint pitch targets shaped by specifications of both functions, but by separate identification of each function. At the same time, the perception can nevertheless follow the temporal progression of sentence production by processing syllable-sized f_0 contours successively.

CRediT authorship contribution statement

Yue Chen: Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Yi Xu:** Writing – review & editing, Supervision, Project administration, Methodology, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported in part by the Graduate Research Scholarships offered by University College London to the author Yue Chen.

Data availability

Data will be made available on request.

References

Abramson, A.S., 1978. Static and dynamic acoustic cues in distinctive tones. Lang. Speech 21 (4), 319–325. https://doi.org/10.1177/002383097802100406.
Alzaidi, M.S.A., Xu, Y., Xu, A., Szreder, M., 2023. Analysis and computational modelling of Emirati Arabic intonation–a preliminary study. J. Phon. 98, 101236. https://doi.org/10.1016/j.wocn.2023.101236.
Alzaidi, M.S., Xu, Y., Xu, A., 2019. Prosodic encoding of focus in Hijazi Arabic. Speech Commun. 106, 127–149. https://doi.org/10.1016/j.specom.2018.12.006.
Ananthakrishnan, S., Narayanan, S., 2005. An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'05), 1, pp. 269–272. https://doi.org/10.1109/ICASSP.2005.1415102.
Anwyl-Irvine, A.L., Massonnié, J., Flitton, A., Kirkham, N., Evershed, J.K., 2020. Gorilla in our midst: an online behavioral experiment builder. Behav. Res. Methods 52 (1), 388–407. https://doi.org/10.3758/s13428-019-01237-x.
Ardali, M.T., Xu, Y., 2012. Phonetic realization of prosodic focus in Persian. Speech Prosody 2012, 326–329. https://doi.org/10.21437/SpeechProsody.2012-83.
Beguš, G., Zhou, A., Zhao, T.C., 2023. Encoding of speech in convolutional layers and the brain stem based on language experience. Sci. Rep. 13 (1), 6480. https://doi.org/10.1038/s41598-023-33384-9.
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. J. Phon. 18 (1), 37–49. https://doi.org/10.1016/S0095-4470(19)30357-2.
Botinis, A., Fourakis, M., Gawronska, B., 1999. Focus identification in English, Greek, and Swedish. In: Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS-14), pp. 1557–1560.
Bruce, G. (1982). Developing the Swedish intonation model. In Lund University, Department of Linguistics Working Papers (Vol. 22, pp. 51–116).
Büring, D., 2006. Focus projection and default prominence. In: Molnár, V., Winkler, S. (Eds.), The Architecture of Focus. De Gruyter Mouton, pp. 321–346. https://doi.org/10.1515/9783110922011.321.
Carlson, K., Dickey, M.W., Frazier, L., Clifton, C., 2009. Information structure expectations in sentence comprehension. Q. J. Exp. Psychol. 62 (1), 114–139. https://doi.org/10.1080/17470210701880171.
Chahal, D., 2003. Phonetic cues to prominence in Lebanese Arabic. In: Proceedings of the 15th International Congress of Phonetic Sciences, pp. 2067–2070.
Chandrasekaran, B., Sampath, P.D., Wong, P.C.M., 2010. Individual variability in cue-weighting and lexical tone learning. J. Acoust. Soc. Am. 128 (1), 456–465. https://doi.org/10.1121/1.3445785.
Chao, Y.R., 1968. A Grammar of Spoken Chinese (2. Print). University of California Press.
Chen, S., Wang, B., Xu, Y., 2009. Closely related languages, different ways of realizing focus. Interspeech 2009, 1007–1010. https://doi.org/10.21437/Interspeech.2009-298.
Chen, Y., 2022. Tone and intonation. In: Huang, C.R., Lin, Y.H., Chen, I.H. (Eds.), The Cambridge Handbook of Chinese Linguistics, 1st ed. Cambridge University Press, pp. 336–360. https://doi.org/10.1017/9781108329019.019.
Chen, Y., Braun, B., 2006. Prosodic realization of information structure categories in standard Chinese. Speech Prosody 2006, 050–051. https://doi.org/10.21437/SpeechProsody.2006-92 paper.
Chen, Y., Gao, Y., Xu, Y., 2022. Computational modelling of tone perception based on direct processing of f0 contours. Brain Sci. 12 (3), 337. https://doi.org/10.3390/brainsci12030337.
Chen, Y., Gussenhoven, C., 2008. Emphasis and tonal implementation in Standard Chinese. J. Phon. 36 (4), 724–746. https://doi.org/10.1016/j.wocn.2008.06.003.
Chen, Y., Xu, Y., 2021. Parallel recognition of Mandarin tones and focus from continuous F0. In: Proceedings of the 1st International Conference on Tone and Intonation (TAI), pp. 171–175. https://doi.org/10.21437/TAI.2021-35.
Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation (arXiv:1406.1078). arXiv. http://arxiv.org/abs/1406.1078.
De Jong, K., 2004. Stress, lexical focus, and segmental focus in English: patterns of variation in vowel duration. J. Phon. 32 (4), 493–516. https://doi.org/10.1016/j.wocn.2004.05.002.
Dohen, M., Loevenbruck, H., 2004. Pre-focal rephrasing, focal enhancement and postfocal deaccentuation in French. Interspeech 2004, 785–788. https://doi.org/10.21437/Interspeech.2004-296.
Face, T.L., 2005. F0 peak height and the perception of sentence type in Castilian Spanish. Rev. Int. Linguist. Iberoam 3 (2), 49–65.
Face, T.L., 2007. The role of intonational cues in the perception of declaratives and absolute interrogatives in Castilian Spanish. Estud. Fon. Exp. 16, 185–225.
Feldman, N.H., Griffiths, T.L., Morgan, J.L., 2009. The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. Psychol. Rev. 116 (4), 752–782. https://doi.org/10.1037/a0017196.
Fernandez, R., Ramabhadran, B., 2010. Discriminative training and unsupervised adaptation for labeling prosodic events with limited training data. Interspeech 2010, 1429–1432. https://doi.org/10.21437/Interspeech.2010-433.
Féry, C., Kügler, F., 2008. Pitch accent scaling on given, new and focused constituents in German. J. Phon. 36 (4), 680–703. https://doi.org/10.1016/j.wocn.2008.05.001.
Francis, A.L., Ciocca, V., Ma, L., Fenn, K., 2008. Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. J. Phon. 36 (2), 268–294. https://doi.org/10.1016/j.wocn.2007.06.005.
Francis, A.L., Ciocca, V., Wong, N.K.Y., Leung, W.H.Y., Chu, P.C.Y., 2006. Extrinsic context affects perceptual normalization of lexical tone. J. Acoust. Soc. Am. 119 (3), 1712–1726.
Fry, D.B., 1958. Experiments in the perception of stress. Lang. Speech 1 (2), 126–152. https://doi.org/10.1177/002383095800100207.
Gandour, J., 1983. Tone perception in Far Eastern languages. J. Phon. 11 (2), 149–175. https://doi.org/10.1016/S0095-4470(19)30813-7.
Gauthier, B., Shi, R., Xu, Y., 2007. Learning phonetic categories by tracking movements. Cognition 103 (1), 80–106. https://doi.org/10.1016/j.cognition.2006.03.002.
Gauthier, B., Shi, R., Xu, Y., 2009. Learning prosodic focus from continuous speech input: a neural network exploration. Lang. Learn. Dev. 5 (2), 94–114. https://doi.org/10.1080/15475440802698524.
Gogoi, P., Dey, A., Lalhminghlui, W., Sarmah, P., Prasanna, S.R.M., 2020. Lexical tone recognition in mizo using acoustic-prosodic features. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 6458–6461. https://aclanthology.org/2020.lrec-1.795.
Gottfried, T.L., Suiter, T.L., 1997. Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones. J. Phon. 25 (2), 207–231. https://doi.org/10.1006/jpho.1997.0042.
Howie, J.M., 1976. Acoustical Studies of Mandarin Vowels and Tones. Cambridge University Press.
Hu, N., Janssen, B., Hanssen, J., Gussenhoven, C., Chen, A., 2020. Automatic analysis of speech prosody in Dutch. Interspeech 2020, 155–159. https://doi.org/10.21437/Interspeech.2020-2142.
Ipek, C., 2011. Phonetic realization of focus with no on-focus pitch range expansion in Turkish. In: Proceedings of the 17th International Congress of Phonetic Sciences, pp. 140–143. https://api.semanticscholar.org/CorpusID:12726947.
Ishihara, S. (2003). Intonation and interface conditions [Ph.D. Dissertation]. Massachusetts Institute of Technology.
Jeon, J.H., Liu, Y., 2009. Automatic prosodic events detection using syllable-based acoustic and syntactic features. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4565–4568. https://doi.org/10.1109/ICASSP.2009.4960646.
Jin, S. (1996). An acoustic study of sentence stress in Mandarin Chinese [Ph.D. Dissertation]. The Ohio State University.
Johnson, K., Sjerps, M.J., 2021. Speaker normalization in speech perception. In: Pardo, J. S., Nygaard, L.C., Remez, R.E., Pisoni, D.B. (Eds.), The Handbook of Speech Perception, 1st ed. Wiley, pp. 145–176. https://doi.org/10.1002/9781119184096.ch6.
Kakouros, S., Räsänen, O., 2016. 3PRO – An unsupervised method for the automatic detection of sentence prominence in speech. Speech Commun. 82, 67–84. https://doi.org/10.1016/j.specom.2016.06.004.
Kakouros, S., Räsänen, O., Alku, P., 2018. Comparison of spectral tilt measures for sentence prominence in speech–effects of dimensionality and adverse noise conditions. Speech Commun. 103, 11–26. https://doi.org/10.1016/j.specom.2018.08.002.
Kakouros, S., Suni, A., Šimko, J., Vainio, M., 2019. Prosodic representations of prominence classification neural networks and autoencoders using bottleneck features. Interspeech 2019, 1946–1950. https://doi.org/10.21437/Interspeech.2019-2984.
Kang, W., Xu, Y., 2024. Tone-syllable synchrony in Mandarin: New evidence and implications. Speech Commun. 163, 103121. https://doi.org/10.1016/j.specom.2024.103121.
Kleinschmidt, D.F., Jaeger, T.F., 2015. Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. Psychol. Rev. 122 (2), 148–203. https://doi.org/10.1037/a0038695.

Kreßel, U.H.G., 1999. Pairwise classification and support vector machines. Advances in Kernel Methods: Support Vector Learning. MIT Press, pp. 255–268.
Kuhl, P.K., 2004. Early language acquisition: cracking the speech code. Nat. Rev. Neurosci. 5 (11), 831–843. https://doi.org/10.1038/nrn1533.
Kuhl, P.K., 2010. Brain mechanisms in early language acquisition. Neuron 67 (5), 713–727. https://doi.org/10.1016/j.neuron.2010.08.038.
Kuhl, P.K., Ramírez, R.R., Bosseler, A., Lin, J.F.L., Imada, T., 2014. Infants' brain responses to speech suggest analysis by synthesis. Proc. Natl. Acad. Sci. 111 (31), 11238–11245. https://doi.org/10.1073/pnas.1410963111.
Ladd, D.R., 2008. Intonational Phonology, 2nd ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511808814.
Lee, A., Chiu, F., & Xu, Y. (2016). Focus perception in Japanese: effects of focus location and accent condition. 060007. 10.1121/2.0000441.
Lee, C.Y., Tao, L., Bond, Z.S., 2008. Identification of acoustically modified Mandarin tones by native listeners. J. Phon. 36 (4), 537–563. https://doi.org/10.1016/j.wocn.2008.01.002.
Lee, Y., Wang, B., Chen, S., Adda-Decker, M., Amelot, A., Nambu, S., Liberman, M., 2015. A crosslinguistic study of prosodic focus. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4754–4758. https://doi.org/10.1109/ICASSP.2015.7178873.
Lee, Y., Xu, Y., 2010. Phonetic realization of contrastive focus in Korean. Speech Prosody 2010, 030–033. https://doi.org/10.21437/SpeechProsody.2010-81 paper.
Lehiste, I., 1970. Suprasegmentals. MIT Press.
Leung, K.K.W., Wang, Y., 2020. Production-perception relationship of Mandarin tones as revealed by critical perceptual cues. J. Acoust. Soc. Am. 147 (4), EL301–EL306. https://doi.org/10.1121/10.0000963.
Levow, G.A., 2005. Context in multi-lingual tone and pitch accent recognition. Interspeech 2005, 1809–1812. https://doi.org/10.21437/Interspeech.2005-552.
Li, Y., Anumanchipalli, G.K., Mohamed, A., Chen, P., Carney, L.H., Lu, J., Wu, J., Chang, E.F., 2023. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. Nat. Neurosci. 26 (12), 2213–2225. https://doi.org/10.1038/s41593-023-01468-4.
Lin, J., Li, W., Gao, Y., Xie, Y., Chen, N.F., Siniscalchi, S.M., Zhang, J., Lee, C.H., 2018. Improving Mandarin tone recognition based on DNN by combining acoustic and articulatory features using extended recognition networks. J. Signal Process. Syst. 90 (7), 1077–1087. https://doi.org/10.1007/s11265-018-1334-2.
Lin, J., Xie, Y., Gao, Y., Zhang, J., 2016. Improving Mandarin tone recognition based on DNN by combining acoustic and articulatory features. In: Proceedings of the 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 1–5. https://doi.org/10.1109/ISCSLP.2016.7918472.
Liu, F. (2009). Intonation systems of Mandarin and English: a functional approach. Dissertations & Theses - Gradworks.
Liu, F., Xu, Y., 2005. Parallel encoding of focus and interrogative meaning in Mandarin intonation. Phonetica 62 (2–4), 70–87. https://doi.org/10.1159/000090090.
Liu, S., Samuel, A.G., 2004. Perception of Mandarin lexical tones when F0 information is neutralized. Lang. Speech 47 (2), 109–138. https://doi.org/10.1177/00238309040470020101.
Liu, Z., Xu, Y., Hsieh, F., 2022. Coarticulation as synchronised CV co-onset–parallel evidence from articulation and acoustics. J. Phon. 90, 101–116. https://doi.org/10.1016/j.wocn.2021.101116.
Massaro, D.W., Cohen, M.M., Tseng, C., 1985. The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese. J. Chin. Linguist. 13 (2), 267–289.
McClelland, J.L., 2013. Integrating probabilistic models of perception and interactive neural networks: a historical and tutorial review. Front. Psychol. 4. https://doi.org/10.3389/fpsyg.2013.00503.
McClelland, J.L., Elman, J.L., 1986. The TRACE model of speech perception. Cogn. Psychol. 18 (1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0.
Meng, H., Chen, C.-T., Chen, Y., Liu, Z., Yi, X., 2023. Mandarin tone production can be learned under perceptual guidance—A machine learning simulation. Proceedings of The 20th International Congress of Phonetic Sciences, pp. 2324–2328.
Mishra, T., Sridhar, V.R., Conkie, A., 2012. Word prominence detection using robust yet simple prosodic features. Interspeech 2012, 1864–1867. https://doi.org/10.21437/Interspeech.2012-408.
Mixdorff, H., 2004. Quantitative tone and intonation modeling across languages. In: Proceedings of the First International Symposium on Tonal Aspects of Languages (TAL 2004), pp. 137–142.
Norris, D., McQueen, J.M., 2008. Shortlist B: a Bayesian model of continuous speech recognition. Psychol. Rev. 115 (2), 357–395. https://doi.org/10.1037/0033-295X.115.2.357.
Norris, D., McQueen, J.M., Cutler, A., 2016. Prediction, Bayesian inference and feedback in speech recognition. Lang. Cogn. Neurosci. 31 (1), 4–18. https://doi.org/10.1080/23273798.2015.1081703.
O'Shaughnessy, D., 1979. Linguistic features in fundamental frequency patterns. J. Phon. 7 (2), 119–145. https://doi.org/10.1016/S0095-4470(19)31045-9.
Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in PyTorch. In: Proceedings of the Conference on Neural Information Processing Systems (NIPS 2017). https://api.semanticscholar.org/CorpusID:40027675.
Patil, U., Kentner, G., Gollrad, A., Kuegler, F.D., Féry, C., Vasishth, S., 2008. Focus, word order and intonation in Hindi. Mind Res. Repos. 1. https://api.semanticscholar.org/CorpusID:15894534.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12 (null), 2825–2830.
Peng, G., Zhang, C., 2015. Tone perception. In: Wang, W.S.Y., Sun, C. (Eds.), The Oxford Handbook of Chinese Linguistics. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199856336.013.0039.
Pierrehumbert, J.B. (1980). The phonology and phonetics of English intonation [Ph.D. Dissertation]. Massachusetts Institute of Technology.
Pierrehumbert, J., Hirschberg, J., 1990. The meaning of intonational contours in the interpretation of discourse. In: Cohen, P.R., Morgan, J., Pollack, M.E. (Eds.), Intentions in Communication. The MIT Press, pp. 271–312. https://doi.org/10.7551/mitpress/3839.003.0016.
Platt, J., 2000. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Smola, A., Bartlett, P., Schölkopf, B., Schuurmans, D. (Eds.), Advances in Large Margin Classifiers. MIT Press.
Prom-on, S., Xu, Y., Thipakorn, B., 2009. Modeling tone and intonation in Mandarin and English as a process of target approximation. J. Acoust. Soc. Am. 125 (1), 405–424. https://doi.org/10.1121/1.3037222.
Qian, Y., Lee, T., Soong, F.K., 2007. Tone recognition in continuous Cantonese speech using supratone models. J. Acoust. Soc. Am. 121 (5), 2936–2945. https://doi.org/10.1121/1.2717413.
Ren, Y., Kim, S.S., Hasegawa-Johnson, M., Cole, J., 2004. Speaker-independent automatic detection of pitch accent. Speech Prosody 2004, 521–524. https://doi.org/10.21437/SpeechProsody.2004-120.
Rosenberg, A., 2010. AutoBI - a tool for automatic toBI annotation. Interspeech 2010, 146–149. https://doi.org/10.21437/Interspeech.2010-71.
Rosenberg, A., Fernandez, R., Ramabhadran, B., 2015. Modeling phrasing and prominence using deep recurrent learning. Interspeech 2015, 3066–3070. https://doi.org/10.21437/Interspeech.2015-623.
Rump, H.H., Collier, R., 1996. Focus conditions and the prominence of pitch-accented syllables. Lang. Speech 39 (1), 1–17. https://doi.org/10.1177/002383099603900101.
Schnall, A., Heckmann, M., 2019. Feature-space SVM adaptation for speaker adapted word prominence detection. Comput. Speech Lang. 53, 198–216. https://doi.org/10.1016/j.csl.2018.06.001.
Shen, X.S., Lin, M., 1991. A Perceptual study of Mandarin tones 2 and 3. Lang. Speech 34 (2), 145–156. https://doi.org/10.1177/002383099103400202.
Shen, X., 沈晓楠, 1989. Interplay of the four citation tones and intonation in Mandarin Chinese /普通话四声与语调的交互关系. J. Chin. Linguist. 17 (1), 61–74.
Shih, C. (1988). Tone and intonation in Mandarin. In N. Clements (Ed.), Working Papers of the Cornell Phonetics Laboratory 3: Stress, tone and intonation (pp. 83–109).
Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., 1992. TOBI: a standard for labeling English prosody. In: Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP 1992), pp. 867–870. https://doi.org/10.21437/ICSLP.1992-260.
Silverman, K.E.A., Pierrehumbert, J.B., 1990. The timing of prenuclear high accents in English. In J. Kingston & M. E. E. Beckman (Eds.). Papers in Laboratory Phonology. Cambridge University Press, pp. 72–106.
Sluijter, A.M.C., Van Heuven, V.J., 1996. Spectral balance as an acoustic correlate of linguistic stress. J. Acoust. Soc. Am. 100 (4), 2471–2485. https://doi.org/10.1121/1.417955.
Stehwien, S., Vu, N.T., 2017. Prosodic event recognition using convolutional neural networks with context information. Interspeech 2017, 2326–2330. https://doi.org/10.21437/Interspeech.2017-1159.
Thorsen, N.G., 1980. A study of perception of sentence intonation–evidence from Danish. J. Acoust. Soc. Am. 67 (3), 1014–1030. https://doi.org/10.1121/1.384069.
Tong, X., Lee, S.M.K., Lee, M.M.L., Burnham, D., 2015. A tale of two features: perception of cantonese lexical tone and English lexical stress in Cantonese-English bilinguals. PLoS ONE 10 (11), e0142896. https://doi.org/10.1371/journal.pone.0142896.
Tupper, P., Leung, K., Wang, Y., Jongman, A., Sereno, J.A., 2020. Characterizing the distinctive acoustic cues of Mandarin tones. J. Acoust. Soc. Am. 147 (4), 2570–2580. https://doi.org/10.1121/10.0001024.
Ullas, S., Bonte, M., Formisano, E., Vroomen, J., 2022. Adaptive plasticity in perceiving speech sounds. In L. L. Holt, J. E. Peelle, A. B. Coffin, A. N. Popper, & R. R. Fay (Eds.). In: Speech Perception, 74. Springer International Publishing, pp. 173–199. https://doi.org/10.1007/978-3-030-81542-4_7.
Van Heuven, V.J., 2018. Acoustic correlates and perceptual cues of word and sentence stress: towards a cross-linguistic perspective. In R. Goedemans, J. Heinz, & H. Van Der Hulst (Eds.). The Study of Word Stress and Accent, 1st ed. Cambridge University Press, pp. 15–59. https://doi.org/10.1017/9781316683101.002.
Van Heuven, V.J., Haan, J., 2002. Temporal distribution of interrogativity markers in Dutch: a perceptual study. In C. Gussenhoven & N. Warner (Eds.). In: Laboratory Phonology, 7. Mouton de Gruyter, pp. 61–86. https://doi.org/10.1515/9783110197105.1.61.
Walsh, M., Schweitzer, K., Schauffler, N., 2013. Exemplar-based pitch accent categorisation using the generalized context model. Interspeech 2013, 258–262. https://doi.org/10.21437/Interspeech.2013-79.
Wang, B., Wang, L., Qadir, T., 2011. Prosodic realization of focus in six languages/dialects in China. In: Proceedings of the 17th International Congress of Phonetic Sciences, pp. 144–147. https://api.semanticscholar.org/CorpusID:11751423.
Wang, B., Xu, Y., 2011. Differential prosodic encoding of topic and focus in sentence-initial position in Mandarin Chinese. J. Phon. 39 (4), 595–611. https://doi.org/10.1016/j.wocn.2011.03.006.
Wang, B., Xu, Y., Ding, Q., 2018. Interactive prosodic marking of focus, boundary and newness in Mandarin. Phonetica 75 (1), 24–56. https://doi.org/10.1159/000453082.
Wang, T., Liu, J., Lee, Y., Lee, Y., 2020. The interaction between tone and prosodic focus in Mandarin Chinese. Lang. Linguist. 語言暨語言學 21 (2), 331–350. https://doi.org/10.1075/lali.00063.wan.

Wang, W.S.Y., 1967. Phonological features of tone. Int. J. Am. Linguist. 33 (2), 93–105. https://doi.org/10.1086/464946.
Wang, W.S.Y., 1972. The many uses of F0. In A. Valdman (Ed.). Papers in Linguistics and Phonetics to the Memory of Pierre Delattre. DE GRUYTER, pp. 487–504. https://doi.org/10.1515/9783110803877-041.
Werker, J.F., Yeung, H.H., 2005. Infant speech perception bootstraps word learning. Trends Cogn. Sci. 9 (11), 519–527. https://doi.org/10.1016/j.tics.2005.09.003.
Whalen, D.H., Xu, Y., 1992. Information for Mandarin tones in the amplitude contour and in brief segments. Phonetica 49 (1), 25–47. https://doi.org/10.1159/000261901.
Wong, P.C.M., Diehl, R.L., 2003. Perceptual normalization for inter- and intratalker variation in cantonese level tones. J. Speech Lang. Hear. Res. 46 (2), 413–421. https://doi.org/10.1044/1092-4388(2003/034.
Wu, T.F., Lin, C.J., Weng, R.C., 2004. Probability estimates for multi-class classification by pairwise coupling. J. Mach. Learn. Res. 5, 975–1005.
Xu, Y., 1994. Production and perception of coarticulated tones. J. Acoust. Soc. Am. 95 (4), 2240–2253. https://doi.org/10.1121/1.408684.
Xu, Y., 1997. Contextual tonal variations in Mandarin. J. Phon. 25 (1), 61–83. https://doi.org/10.1006/jpho.1996.0034.
Xu, Y., 1999. Effects of tone and focus on the formation and alignment of f0contours. J. Phon. 27 (1), 55–105. https://doi.org/10.1006/jpho.1999.0086.
Xu, Y., 2005. Speech melody as articulatorily implemented communicative functions. Speech Commun. 46 (3–4), 220–251. https://doi.org/10.1016/j.specom.2005.02.014.
Xu, Y., 2013. ProsodyPro–a tool for large-scale systematic prosody analysis. In: Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013), pp. 7–10.
Xu, Y., 2015. Intonation in Chinese. In: Wang, W.S.Y., Sun, C. (Eds.), The Oxford Handbook of Chinese Linguistics. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199856336.013.0012.
Xu, Y. (2020). Syllable is a synchronization mechanism that makes human speech possible [Preprint]. PsyArXiv. 10.31234/osf.io/9v4hr.
Xu, Y., Chen, S., Wang, B., 2012. Prosodic focus with and without post-focus compression: a typological divide within the same language family? Linguist. Rev. 29 (1). https://doi.org/10.1515/tlr-2012-0006.
Xu, Y., Prom-On, S., 2014. Toward invariant functional representations of variable surface fundamental frequency contours: synthesizing speech melody via model-based stochastic learning. Speech Commun. 57, 181–208. https://doi.org/10.1016/j.specom.2013.09.013.
Xu, Y., Wang, Q.E., 2001. Pitch targets and their realization: evidence from Mandarin Chinese. Speech Commun. 33 (4), 319–337. https://doi.org/10.1016/S0167-6393(00)00063-7.
Xu, Y., Xu, C.X., 2005. Phonetic realization of focus in English declarative intonation. J. Phon. 33 (2), 159–197. https://doi.org/10.1016/j.wocn.2004.11.001.
Xu, Y., Xu, C.X., Sun, X., 2004. On the temporal domain of focus. Speech Prosody 2004, 81–84. https://doi.org/10.21437/SpeechProsody.2004-19.
Yan, J., Meng, Q., Tian, L., Wang, X., Liu, J., Li, M., Zeng, M., Xu, H., 2023. A Mandarin tone recognition algorithm based on random forest and feature fusion. Mathematics 11 (8), 1879. https://doi.org/10.3390/math11081879.
Yan, M., Calhoun, S., 2020. Rejecting false alternatives in Chinese and English: the interaction of prosody, clefting, and default focus position. Lab. Phonol. J. Assoc. Lab. Phonol. 11 (1), 17. https://doi.org/10.5334/labphon.255.
Yip, M., 2002. Tone, 1st ed. Cambridge University Press. https://doi.org/10.1017/CBO9781139164559.
Yu, K.M., 2017. The role of time in phonetic spaces: temporal resolution in Cantonese tone perception. J. Phon. 65, 126–144. https://doi.org/10.1016/j.wocn.2017.06.004.
Yu, K.M., Lam, H.W., 2014. The role of creaky voice in Cantonese tonal perception. J. Acoust. Soc. Am. 136 (3), 1320–1333. https://doi.org/10.1121/1.4887462.
Yuan, J., 2011. Perception of intonation in Mandarin Chinese. J. Acoust. Soc. Am. 130 (6), 4063–4069. https://doi.org/10.1121/1.3651818.
Zhang, C., Chen, S., 2016. Toward an integrative model of talker normalization. J. Exp. Psychol. Hum. Percept. Perform. 42 (8), 1252–1268. https://doi.org/10.1037/xhp0000216.
Zhang, C., Shao, J., Chen, S., 2018. Impaired perceptual normalization of lexical tones in Cantonese-speaking congenital amusics. J. Acoust. Soc. Am. 144 (2), 634–647. https://doi.org/10.1121/1.5049147.
Zhang, H., Wiener, S., Holt, L.L., 2022. Adjustment of cue weighting in speech by speakers and listeners: evidence from amplitude and duration modifications of Mandarin Chinese tone. J. Acoust. Soc. Am. 151 (2), 992–1005. https://doi.org/10.1121/10.0009378.
Zhang, J., 2022. Tonal processes defined as tone Sandhi. In: Huang, C.R., Lin, Y.H., Chen, I.H. (Eds.), The Cambridge Handbook of Chinese Linguistics, 1st ed. Cambridge University Press, pp. 291–312. https://doi.org/10.1017/9781108329019.017.
Zhang, J., Hirose, K., 2004. Tone nucleus modeling for Chinese lexical tone recognition. Speech Commun. 42 (3–4), 447–466. https://doi.org/10.1016/j.specom.2004.01.001.
Zhu, X., Wang, C., 2015. Tone. In: Wang, W.S.Y., Sun, C. (Eds.), The Oxford Handbook of Chinese Linguistics. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199856336.013.0011.