PERSPECTIVES SIG 19

Research Note

The Bubble Noise Technique for Speech Perception Research

Michael I. Mandel,[a,b] Vikas Grover,[c] Mengxuan Zhao,[b] Jiyoung Choi,[d] and Valerie L. Shafer[b]

Purpose: The "bubble noise" technique has recently been introduced as a method to identify the regions in time–frequency maps (i.e., spectrograms) of speech that are especially important for listeners in speech recognition. This technique identifies regions of "importance" that are specific to the speech stimulus and the listener, thus permitting these regions to be compared across different listener groups. For example, in cross-linguistic and second-language (L2) speech perception, this method identifies differences in regions of importance in accomplishing decisions of phoneme category membership. This research note describes the application of bubble noise to the study of language learning for 3 different language pairs: Hindi English bilinguals' perception of the /v/–/w/ contrast in American English, native English speakers' perception of the tense/lax contrast for Korean fricatives and affricates, and native English speakers' perception of Mandarin lexical tone.
Conclusion: We demonstrate that this technique provides insight on what information in the speech signal is important for native/first-language listeners compared to nonnative/L2 listeners. Furthermore, the method can be used to examine whether L2 speech perception training is effective in bringing the listener's attention to the important cues.

Studying how word and sentence meaning is recovered from the speech signal is an enduring pursuit because this process is complex, both due to the complex nature of the speech signal and to the complexity of the listener. The goal of this report is to introduce a novel method, called "bubble noise," that can be used in this pursuit. We aim to familiarize readers with this technique who might be interested in applying it to their own research questions in speech perception. Note that MATLAB code for generating the noise stimuli, designing and performing the task, and analyzing the results is freely available on GitHub (2019).

The specific goal of this research note is to offer a tutorial in a newly introduced bubble noise method in the area of nonnative speech perception. We will mainly focus on methodological considerations. This research note is organized as follows: In the Introduction to Speech Perception Research section, we briefly define and describe speech perception research. In The Bubble Noise Technique section, we describe the bubble noise procedure and basic design decisions in using this method. In the Example Applications in L2 Learning/Perception section, we present three sample experiments of cross-linguistic speech perception to illustrate the method. We finish by discussing the strengths and weaknesses of this method.

Introduction to Speech Perception Research

We assume that most of our readers have some background in speech perception research, but briefly describe the field here (novice readers are referred to Diehl, Lotto, & Holt, 2004; Samuel, 2011, for reviews of the field). Speech perception is the act of recovering the phonological identity of an acoustic signal that typically is produced by the human vocal apparatus (but can be applied to synthetic signals, such as sine-wave speech or nonhuman vocal productions, e.g., from parrots; Strange, 2011; Strange & Shafer, 2008). We typically view phonological recovery in terms of phonological segments (e.g., /b/, /d/, /a/, /s/). In English, these roughly correspond to our alphabetic orthographic system. However, the speech signal corresponding to a word or

[a]Brooklyn College, Brooklyn, NY
[b]CUNY Graduate Center, New York, NY
[c]New York Medical College, Valhalla, NY
[d]Lehman College CUNY, New York, NY
Correspondence to Michael Mandel: mim@sci.brooklyn.cuny.edu

Editor: Angela H Ciccia
Received August 27, 2019
Accepted September 5, 2019
https://doi.org/10.1044/2019_PERS-19-00058

Disclosures
Financial: Michael I. Mandel has no relevant financial interests to disclose. Vikas Grover has no relevant financial interests to disclose. Mengxuan Zhao has no relevant financial interests to disclose. Jiyoung Choi has no relevant financial interests to disclose. Valerie L. Shafer has no relevant financial interests to disclose. Nonfinancial: Michael I. Mandel has no relevant nonfinancial interests to disclose. Vikas Grover has no relevant nonfinancial interests to disclose. Mengxuan Zhao has no relevant nonfinancial interests to disclose. Jiyoung Choi has no relevant nonfinancial interests to disclose. Valerie L. Shafer has no relevant nonfinancial interests to disclose.

utterance does not neatly divide into these segment-sized units. Production of speech includes transitions from one vocal shape to another, and these transitions are reflected in the acoustic signal. In addition, variation in the vocal apparatus across speakers, and variations in the productions of a speaker (often due to prosodic-level decisions that vary fundamental frequency [F0] and rhythm) result in variations in the acoustic information corresponding to a phonological segment. As a consequence, recovery of the phonological identity is by no means trivial. A large literature has focused on understanding how humans identify phonological units (Diehl et al., 2004; Liberman, 1982; Samuel, 2011). The principle methods for studying speech perception have been to examine identification ("What speech sound did you hear?") and/or discrimination ("Are two speech sounds different?"). Many variations of identification and discrimination tasks have been used (Strange, 2011). We will not review this literature here because our goal is merely to illustrate a novel method that can be used to examine speech perception. However, in the Discussion Relative to Other Methods section, we will discuss the strengths and weaknesses of the bubble noise technique in comparison to some of these commonly used methods.

The Bubble Noise Technique

The bubble noise technique is designed to identify relevant (important) regions of the speech signal that are used in speech perception (Mandel et al., 2016). This technique mixes a small set of speech tokens (e.g., tokens of [ba], [fa], [va], and [wa]) with specially designed interference that includes "holes" or "bubbles" in noise. These bubbles allow glimpses of the target signal (i.e., the token). Participants are asked to identify the token from a closed set (e.g., press the key labeled "b," "f," "v," or "w" that corresponds to what you heard), that is, an *N*-alternative forced-choice design (although others are possible). This procedure is repeated several hundred times with a fairly small set of tokens (typically two to six) in order to explore many different random combinations of speech cues revealed through the bubble noise for each token. An analysis is then performed on each token type (e.g., [wa] tokens) to identify correlations between the correct identification of the token and the audibility of various portions of the speech localized in both time and frequency. A high correlation between bubble time–frequency location and response accuracy indicates the importance of this location in identifying the phonological target. In other words, this region contains a cue that the listener is using to perform this identification, that is, an important cue. These regions are illustrated as importance maps, visualizations of importance as a function of time and frequency. The method was inspired by the "Bubbles" technique in vision research (Gosselin & Schyns, 2001), which has been used to identify various visual features used by viewers in making visual classifications.

The noise used in these experiments has been designed to facilitate the analysis of importance via correlation. This noise is quite "loud" compared to the speech target (typically around 20 dB more intense). As stated above, the holes (or bubbles) in the noise are placed randomly in time and frequency, revealing a different random combination of speech cues in each mixture. Figure 1 shows spectrograms of bubble noise (1a), the bubble noise masking a speech token (1b), and two of the target speech tokens used in the study (1c and 1d). The red regions of these spectrograms show higher intensity noise, and the elliptical, blue "bubbles" are the holes (decreased noise intensity) revealing localized "glimpses" of the speech signal (Cooke et al., 2006). Notice how the speech information (shown as the spectrogram in Figure 1c) inside of the bubbles is nearly unaffected by the noise, but the speech information outside of the bubbles is completely obscured. Thus, for a given mixture involving bubble noise, the location and extent of the bubbles indicate the portions of the speech that are audible to the listener and, in subsequent analysis, thus can be used to identify those regions whose audibility correlates significantly with correct identification.

The remainder of this section provides an overview of the technique, including details that are important to keep in mind when one is designing and executing a study using it. Three such studies are then described in the Example Applications in L2 Learning/Perception section to illustrate the technique and the effect of these details on their implementation.

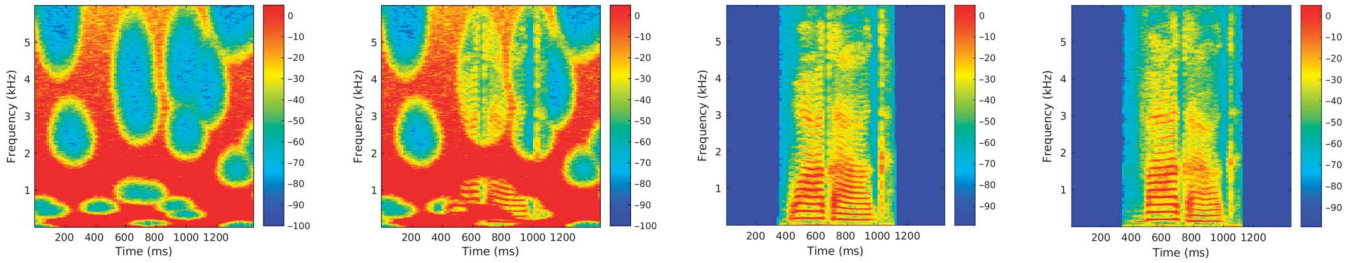
Noise Substrate

The specifics of the noise process are as follows. The noise substrate, from which the bubbles are removed, consists of speech-shaped noise that is much greater in intensity than the speech signal (e.g., –20 dB signal-to-noise ratio [SNR]). This SNR is set so that, without any bubbles, the speech is completely unintelligible and subjects listening to mixtures of it guess the correct choice at chance levels (chance is determined by the number of alternative choices). Once this SNR is established, bubbles are cut out of the noise. Bubble locations are uniformly distributed at random in both the time and frequency dimensions, where time is measured in seconds and frequency is measured on a perceptually motivated scale, which is logarithmic in nature: the Equivalent Rectangular Bandwidth for Normal-hearing listeners (ERB_N) scale (Glasberg & Moore, 1990). The bubble dimensions are also uniform in time and in ERB_N frequency, leading to "taller-looking" bubbles at high frequencies and "shorter-looking" bubbles at low frequencies in our linear-frequency visualizations. Figure 1 illustrates this pattern. The bubbles have a half-amplitude (i.e., noise is suppressed by 3 dB in this region) "width" of 90 ms at their widest and a half-amplitude "height" of 1 ERB_N at their tallest, the smallest values that would avoid introducing audible artifacts. At the center of each bubble, the noise is suppressed by 80 dB, with that suppression falling off quadratically in linear units from that point.

Design Features

As with most speech perception designs, the bubble noise technique requires multiple trials of the same target

Figure 1. Example task: Discriminating the word "vɑgɑg" from the word "wɑgɑg" along with one instance of bubble noise and its mixture with one of the utterances under evaluation ("wɑgɑg"). The word for this trial was correctly identified as "wɑgɑg" rather than "vɑgɑg" by a native English listener. Color represents intensity in decibels, with greater intensity in the spectrogram in red, whereas low power is in blue (seen as the bubbles).



category. Thus, it is essential to consider how repetition of stimulus tokens of categories might modulate the identification response. Similar to other studies of speech perception, it is crucial to design the stimuli to preclude listeners from making decisions about the token identity that are unrelated to the information of interest. Specifically, if the goal of the study is to understand what cues are used for identifying /wa/ versus /va/, it is crucial to make sure that other elements (e.g., F0, intensity, vowel quality) are not correlated with the target information. Including some variation in the nontarget information, however, may be useful. In particular, naturally produced speech cannot easily be controlled for the nontarget properties in a speech token. In addition, the researcher may want to allow some natural variation for ecological validity (Strange, 2011). Thus, the use of several tokens of the same category that show similar variability for the different types may be desirable. For example, Grover (2016) selected four different tokens of each phoneme type (/v/ and /w/) to test perception of American English (AE) /v/ and /w/ by Hindi speakers, using traditional discrimination (AXB categorical discrimination) and identification tasks. These tokens were selected so that they had consistent pitch contour, timing, and loudness, but also included some natural variation. In the next section, one of the studies that we present used these /v/ and /w/ contrasts (in the nonsense words "vɑgɑg" and "wɑgɑg"; see Figure 1 for spectrograms of two of these tokens).

In addition, experiments using longer speech sequences need to consider factors such as transitional probabilities across syllables and words, as well as semantic and syntactic relationships. If the focus is on speech perception, rather than higher level linguistic properties, then it is necessary to ensure that these factors are not correlated with the target speech information (although, this technique could also be used to examine how these higher level cues modulate speech perception).

Adaptive Experimental Design

The difficulty of the listener's task can be adjusted by changing the number of bubbles revealed in a given mixture. More bubbles make the task easier, and fewer bubbles make the task more difficult. One improvement introduced in the bubble noise technique beyond what is described by Mandel et al. (2016) is the use of the weighted up-down procedure (Kaernbach, 1991), which is an adaptive design that adjusts the difficulty for each listener (see Levitt, 1971, for the first study using an adaptive design). Using this procedure, the number of bubbles delivered per second is decreased by a small amount whenever the listener correctly identifies a stimulus and is increased by a small amount otherwise. In particular, when the listener selects the correct response, the bubbles-per-second (BPS) level is divided by 1.02 (resulting in fewer noise-free regions). When the listener selects an incorrect response, this level is multiplied by $1.02^{\alpha/(1-\alpha)}$, where α is a desired proportion of correct responses from the listener. Kaernbach (1991) shows that updating the difficulty in this way will eventually converge to the desired proportion of responses, α. The base factor of 1.02 controls the speed of convergence; increasing it will result in faster convergence but will decrease the stability of the convergence. Our target accuracy, α, is $0.5 \times (1 + 1/K)$ for K choices, which is halfway between the chance level and perfect accuracy.

It is possible to adapt the degree of difficulty across all stimuli or, alternatively, for each stimulus individually. Using the latter approach permits a better fit to situations in which certain stimuli are truly easier or harder to identify than others. However, such differences in noise level could be noticeable to subjects, and this noise difference could be used as a cue for specific tokens, confounding the results. Our software package defaults to adapting parameters for individual stimuli, but it is advisable to switch to global adaptation if piloting of a study indicates too great a noise level difference across stimulus tokens.

We typically start listeners at a difficulty of around 20 bubbles per second and allow the adaptive design to adjust this as necessary. Our analysis requires approximately 200 mixtures per utterance, so over the course of these $200 \times K$ trials, this adaptive design tends to converge well. Once the adaptation procedure converges, the accuracy of all listeners at identifying all of the utterances should be

close to the desired level, **α**; thus, it is not useful in comparing the abilities of different listeners to recognize these words. Instead, the final BPS level can be used in this way, with lower BPS levels indicating higher proficiency and higher BPS levels indicating lower proficiency of speech recognition in noise. The listening test design is also blocked so that, for K utterances, each block of $4K$ mixtures will include four copies of each of the K utterances. Although the adaptive design allows flexibility in stimulus construction for listeners of different abilities, it necessitates mixtures being generated on-the-fly, resulting in only one listener hearing each mixture and only hearing it once.

Analysis by Correlation

Once the intelligibility of a sufficient number of mixtures has been measured, the correspondence between audibility at each spectrogram point and intelligibility can be computed. We do this by first computing the spectrogram of the noise added to each mixture, from which we calculate the attenuation of the noise at each point in this spectrogram. Such points are also known as time–frequency points as they represent locations on the time–frequency plane of the spectrogram. This attenuation is scaled to the range of 0–1, where 0 indicates that the noise is fully attenuated and 1 indicates that the noise is not attenuated at all. This attenuation is measured at each time–frequency point and is denoted $M(f,t)$. Following Calandruccio and Doherty (2007), we utilize the point biserial correlation; this is simply a correlation between a continuous variable ($M_i(f,t)$, the noise attenuation amount at frequency f and time t for mixture i) and a dichotomous variable (y_i, 1 if the listener correctly identified the utterance in mixture i and 0 otherwise). This test is performed at every time–frequency point. For example, for a 2-s mixture, this is approximately 100,000 points. The significance of this correlation at each time–frequency point can be measured using a one-way analysis of variance (ANOVA) with two levels, leading to a matrix of p-values, $p(f,t)$. These p-values are visualized by adjusting the color value (in hue–saturation–value color space) of spectrogram points to be $v(f,t) = 0.5\,(1 + \exp(-p(f,t) \backslash \theta))$ so that points with p-values much smaller than θ are shown at full value (bright), whereas points with p-values much larger than θ are shown at half value (dark). See Figure 2 for an example of this visualization process.

Performing this many statistical tests in parallel will obviously lead to false detections. We control the false detection rate of these tests using the method of Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) to set the threshold θ. In particular, when performing many statistical tests in parallel, as we are, this method corrects the results so that only a fixed proportion of identified significant results may be spurious. In our case, we use a proportion of 5%, for example, if 1,000 important time–frequency points are identified out of 100,000, then, at most, 50 of those identifications are likely to be spurious. The original method of Benjamini and Hochberg is appropriate when the tests are positively correlated or uncorrelated; the method of Benjamini and Yekutieli is appropriate if the tests are negatively correlated.

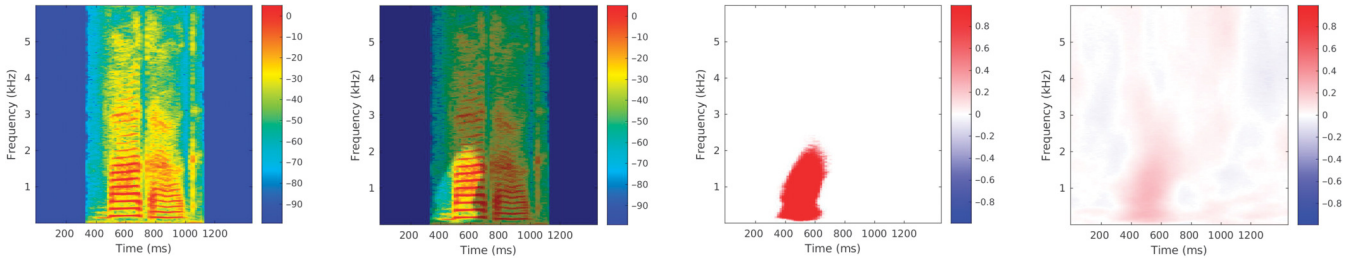
When analyzing results from multiple listeners, we typically perform this analysis over the pooled responses from all listeners. Mandel et al. (2016) measured the consistency between listeners on a six-way forced-choice task with five native English listeners on identical mixtures. They found a moderate amount of across-subject consistency in responses to specific mixtures, but a large amount of consistency in the derived importance maps. There was also a large amount of consistency between the importance maps derived for the individual listeners and maps derived for all of the listeners together. Pooling responses across subjects is warranted when all subjects are following approximately the same listening strategy in a task using the same cues. When such pooling is not warranted, there are other ways to compare the analyses of individual subjects, although these are beyond the scope of this research note. In the experiments below, analyses are performed using this across-subject pooling, as we believe the language backgrounds of each subject group are sufficiently similar for the given tasks that they are using the same strategy.

Example Applications in Second-Language Learning/Perception

This section describes three experiments using the bubble noise technique to study second-language (L2) learning and perception. The purpose of describing them is to provide three illustrative examples of the types of questions that can be addressed using the bubble noise technique. In particular, the experiment in the Hindi /v/–/w/ Contrast section investigates a difficulty in the perception of a contrast in English for a specific population of L2 learners, the experiment in the Korean Tense/Lax Distinction for Fricatives and Affricates section investigates a difficulty in the perception of a Korean contrast for native speakers of English, and the experiment in the Mandarin Tone "Coarticulation" section investigates the perception of coarticulation of lexical tones by native speakers of nontonal languages. All experiments were performed using the authors' Auditory Bubbles toolbox mentioned above, written in the MATLAB programming language. This includes noise generation and mixing, mixture presentation and adaptation, and all analyses. Unless otherwise mentioned, the experiments use the methods and procedures described in The Bubble Noise Technique section, for example, using a false detection rate of 0.05.

One complication in applying the bubble noise technique to L2 perception is that the correlational analysis described above can only be computed if listeners are able to correctly identify some stimulus tokens in the presence of bubble noise. The adaptive design can only achieve this goal if listeners are able to correctly identify each utterance in the absence of noise. For example, consider the case where an L2 learner performs at chance on discriminating tokens of an L2 phoneme contrast that is not in the listeners'

Figure 2. Example correlational analysis of bubble noise results for a native English listener on "vɑgɑg": (a) the spectrogram of the stimulus token; (b) the region allowing for correct identification (highlighted by the brightened color); (c) the time–frequency points that are significantly correlated with correct responses, after correcting for false discovery rate; (d) the correlations themselves. Note that color bars for importance maps are shown at their full brightness. Unimportant points at a given intensity have the same hue and saturation, but a lower brightness.



first language (L1). The adaptive procedure will increase the BPS to a rate where there is effectively no noise masking the signal, but the listener will still be performing at chance. Thus, in the study of naïve listeners, or L2 learners, it is first necessary to train the participants on discrimination of the noise-free tokens. This is done using a forced-choice task with feedback on correctness for each response. Participants repeat this task until they are able to correctly identify each utterance in at least 90% of trials. Most nonnative subjects were able to do so with 5–10 min of training. After familiarizing subjects with the clean utterances, we run a second forced-choice task with feedback to familiarize them with the bubble noise. Then, participants receive the main experimental task, which does not include feedback.

Although the ability of naïve listeners to learn to make nonnative distinctions so quickly may be surprising, note that this is only for a small set of tokens from a single talker. Considerable research has shown that naïve participants can be trained to perform well on a small set of examples of a nonnative contrast (e.g., tokens with /l/ vs. /r/ by Japanese listeners), but that this training will not generalize to novel tokens and different speakers without further training on this larger set of stimuli (e.g., Bradlow, 2008). What is useful regarding the bubble noise approach is that it can identify in what way these nonnative speakers (or L2 learners) are succeeding in the task and how their performance changes with training. In other words, the method can identify when they are focusing on the incorrect regions of the speech signal.

Hindi /v/–/w/ Contrast

In English, but not Hindi, /v/ and /w/ are phonologically contrastive. Hindi includes a labiodental approximant /ʋ/ that is phonetically similar to /v/ and a /w/ and may be substituted for /v/ and /w/ in Hindi English. Hindi speakers of English perceive and produce the English /v/–/w/ contrast less accurately (near chance identification) than AE speakers (Grover, Shafer, Campanelli, Whalen, & Levy, 2019); the source of this misperception is currently unclear, but production data provide some clues. Grover (2016) reports an experiment in which Hindi speakers' productions of AE /v/ and /w/ were rated by AE listeners using a 7-point rating scale. Only a small proportion of the /v/ and /w/ tokens were rated as clear exemplars of either /v/ or /w/. Preliminary acoustic analysis of Hindi speakers' productions of AE /v/ and /w/ suggests differences in the second formant (F2) onset and F2 slope (in the range of 500–1500 Hz; Grover, Shafer, Whalen, Levy, & Kakadelis, 2018). Formant frequencies are related to the shape of the vocal tract. The F2 is specifically related to lip rounding and tongue advancement/backing. Preliminary analysis indicated steeper F2 slopes for AE than Hindi speakers in producing /w/ targets. Hindi speakers showed little difference in these values when producing /v/ and /w/ targets (target prompts included spoken and written forms). This suggests that Hindi speakers' productions of /w/ as compared to the AE speakers' productions lack lip rounding before transitioning into a vowel in nonsense words such as "wɑgɑg." These production data provide clues as to how Hindi listeners might misperceive English /v/ and /w/, but direct measures of perception are necessary to be certain.

We used the bubble noise technique to examine what portion of the speech signal is used by Hindi listeners to identify English /v/ and /w/ tokens. English is an L2 for the tested participants (note, however, that many Hindi speakers are proficient in three or more languages). Our first hypothesis was that Hindi listeners would require more bubbles per second to accurately identify /v/ and /w/ as compared to the English listeners, indicating Hindi speakers' need for more acoustic information to distinguish these two speech sounds. Our second hypothesis was that Hindi listeners would demonstrate less accurate identification of /v/ and /w/ than English listeners. Our third hypothesis was that the findings from the bubble noise design would point to the same time–frequency regions as found using acoustic analysis that correspond to the regions for F2 onset, indicating the lack of awareness in Hindi listeners for the presence or absence of lip rounding in the identification of /w/ and /v/ and the need for targeted training to identify these speech sounds.

Method

Participants. Six AE listeners with normal hearing and six Hindi speakers of English (age range: 26–46 years) identified the target consonant in CVCVC nonsense words. All Hindi speakers learned English in India and arrived in the United States after 18 years of age. An additional Hindi speaker received more substantial training on this contrast and is included to demonstrate how this method can be used to examine the effects of training.

Stimuli. This experiment presented 800 tokens of four words produced by an AE speaker in CVCVC nonsense words; /vɑgɑg/, /wɑgɑg/, /bɑgɑg/, and /fɑgɑg/ were used in this study (see Grover, 2016). The target speech sounds were /v/ and /w/, and the control speech sounds were /b/ and /f/. The phonemes /b/ and /f/ served as control consonants because they are contrastive in Hindi and should be easily categorized and discriminated.

The stimuli used were adapted from the study of Grover (2016). One monolingual female speaker of AE (from the New York State region) recorded naturally produced English nonsense word stimuli for the experimental study. Multiple tokens of nonsense words were recorded on a Dell computer with a Turtle Beach Montego II sound card, using Shure (Model SM 10A) head-worn microphone in a sound-shielded booth and digitized at 22050 Hz using a Sound Forge (Version 4.5). From this large set, final stimuli were selected, which were similar in F0 (range: 189–220 Hz; mean: 192 Hz). Table 1 provides the mean syllable duration and standard deviation for each word type. The final selected word forms were normalized for intensity by root-mean-square using Adobe Audition (Version 6). Stimuli were labeled and verified by three AE listeners who did not participate in stimulus production or in the experiment, in order to ensure accuracy in the production of the intended consonant. Any items that were not identified with 100% accuracy were removed and rerecorded.

Procedure. The stimulus words were presented from a laptop (MacBook Pro) by means of MATLAB software and a high-quality Razer Kraken 7.1 Chroma headset at a range of 75–90 dB SPL. Prior to the study, stimulus intensity was calibrated using a sound-level meter (Larson Davis 800B Precision Integrating Sound Level Meter). Participants were allowed to adjust the volume to a comfortable level at the beginning of the experiment within this range. Participants were asked to complete a four-alternative forced-choice identification task, where they had to identify the sound they heard from four choices they were provided. This experiment presented approximately 200 tokens of each stimulus. The difficulty level of the task was adjusted in the "bubble noise" design, as described above. The adaptive design targeted a listener accuracy level of 62.5%, the target level for a four-alternative forced-choice task. The BPS parameter was controlled globally across all stimuli.

Table 1. Nonsense word forms and syllable duration in milliseconds by condition.

Consonant	Initial position	Duration: *M* (*SD*)
/v/	ˈvagag	178.5 (16.8)
/w/	ˈwagag	198.1 (14.5)
/f/	ˈfagag	152.5 (8.0)
/b/	ˈbagag	180.2 (9.3)

Results

The results are shown in Figure 3. The left column shows spectrograms of the clean speech for /vɑgɑg/ and /wɑgɑg/, and the center column shows the importance maps of individual spectrogram points to AE listeners and Hindi listeners for both words overlaid on the spectrograms. The right column shows the raw correlation values of important regions for AE and Hindi listeners.

These results show that native English speakers are focusing narrowly on the onset of the F2 region to make the decision (i.e., the first 100–200 ms after onset and between 500 and 2500 Hz). Hindi speakers are focusing on the first 200 ms, but in a lower spectral range (below 2000 Hz) for /wɑgɑg/ tokens, but they show no region of importance for /vɑgɑg/, suggesting variability across the six listeners. The trained Hindi speaker, who had engaged in multiple repetitions of the stimuli for training, showed similar regions to the English speakers for identifying /v/ and /w/ (see Figure 4). This finding suggests success in training of this contrast, although it will be necessary to test whether this performance extends to novel tokens.

Final BPS levels. The ability of the listeners to perform this identification task in noise can be evaluated through the final BPS level. In these preliminary experiments, the English speakers achieved the target accuracy with 24.9 BPS. The Hindi speakers, in contrast, ended the experiment at an average of 65.2 BPS, but still did not achieve the target accuracy on some words (they achieved around 40% accuracy on the /v/–/w/ contrast). The trained Hindi speaker achieved the target accuracy at 20.5 BPS.

Not surprisingly, for AE listeners, the regions identified by the bubble noise experiment matched those identified in the preliminary acoustic analysis as production data. Of particular value was the information provided regarding the trained Hindi listener. This participant showed a similar pattern of correlation between signal audibility and identification of the /v/–/w/ contrast to AE listeners. Based on these results, the bubble noise design appears to be a promising method for investigating the specific regions that should be the focus of training and whether training has been effective in highlighting these regions.

Korean Tense/Lax Distinction for Fricatives and Affricates

An investigation of L2 learners of Korean serves to illustrate how the bubble noise technique can elucidate L2 learning of a difficult laryngeal distinction. Korean has a distinctive three-way laryngeal distinction in stops and affricates and a two-way distinction in fricatives. The three-way laryngeal contrast in stops and affricates consists of

Figure 3. Importance maps for native English speakers and native Hindi speakers on the words /vɑgɑg/ and /wɑgɑg/. The left column shows the original spectrogram, the middle column shows the spectrogram overlaid with important regions, and the right column shows the correlation between audibility and correct identification of the word.

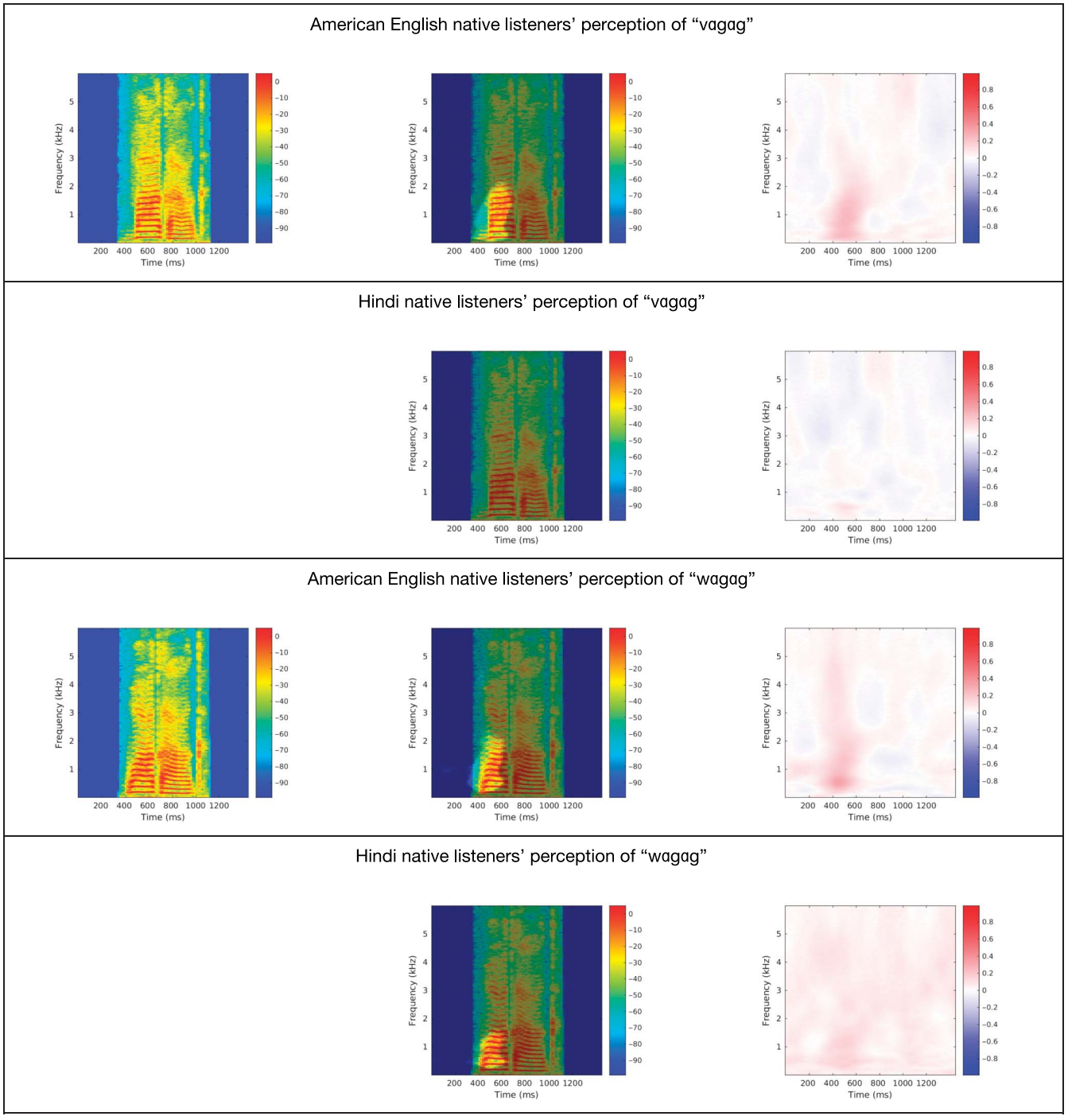
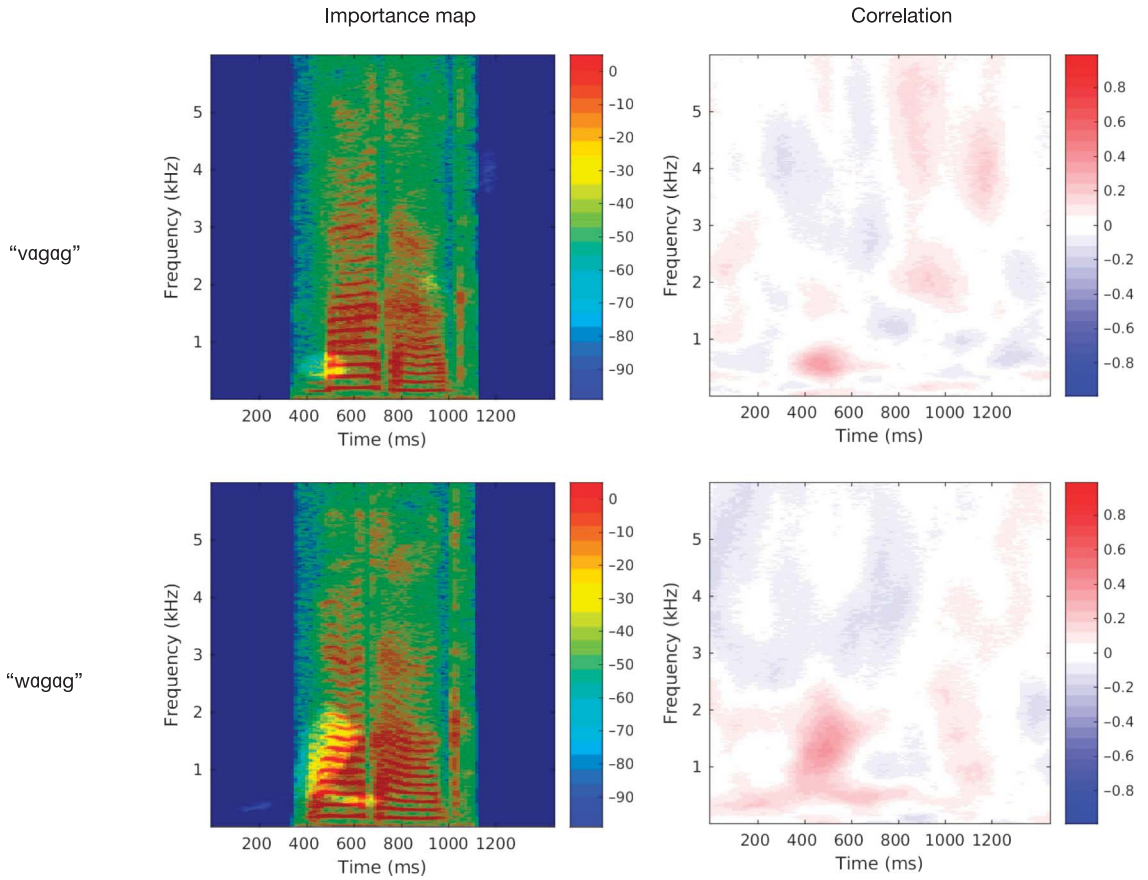


Figure 4. The top row shows the spectrogram for the trained Hindi speaker's responses for "vɑgɑg," and the bottom row shows the spectrogram for the trained Hindi speaker's responses for "wɑgɑg."



aspirated, lax, and tense (Cheon & Anderson, 2008), also labeled as aspirated, lenis, and fortis (Cho, 1995; Cho & Keating, 2001; Dart, 1987). The nature of the dental–alveolar fricative contrast is currently debated in the literature, with one group suggesting these should be described in terms of lax/tense (lenis/fortix) and the other suggesting they should be described in terms of aspiration (Chang, 2007; Chang, 2013; S. Kim, 2000; Park, 1999; Shin, 2001; Yoon, 1999).

Native speakers of Korean use several cues to produce and perceive these distinctions. In initial position, these include voice onset time (VOT), fundamental frequency of the following vowel, voice quality of the following vowel, and the ratio of energy in the first and second harmonics (e.g., M.-R. Kim, Beddor, & Horrocks, 2002). In English, the fortis/lenis distinction for initial affricates is primarily made using VOT. Thus, we hypothesize that native English speakers will primarily focus on the VOT of the Korean affricates and sibilants and largely ignore the information in the subsequent vowel. The bubble noise technique can show the temporal extent of the cues used by native English speakers and whether it includes portions of the subsequent vowel or not.

Method

Participants. Seventeen subjects participated in the study: 12 native Korean speakers (age range: 19–38 years) and five L1 AE speakers (age range: 17–24 years) whose Korean proficiency was at a novice level and who were enrolled in elementary Korean courses at Queens College. Their experience learning Korean ranged from 4 to 9 months, including both informal and formal education. All participants provided informed consent. The study was conducted under the supervision of The City University of New York Institutional Review Board.

Stimuli. Five Korean words were used: /tɕɑdɑ/ (to sleep), /tɕ^{h}ɑdɑ/ (to kick), /tɕ*ɑdɑ/ (to squeeze), /s^{h}ɑdɑ/ (to buy), and /s*ɑdɑ/ (cheap) with a CVCVC form. Tense is indicated by [*], lax is indicated by no superscript, and aspiration is indicated by [^{h}]. The stimuli were recorded by a native female speaker of Korean in a sound-treated room with a Shure SM48 microphone at a calibrated distance. All stimuli were target speech words. The speaker repeated the words five times with various speech styles, including different pitch, intensity, and speed. One repetition of each word was selected such that all selected recordings approximately matched in pitch, duration, and intensity.

The intensity of the stimuli presentation was adjusted by the subjects to a comfortable level at the beginning of the experiment.

Design and procedure. Romanized letters "j" = /tɕ/, "ch" = /tɕʰ/, "jj" = /tɕ*/, "s" = /sʰ/, and "ss" = /s*/ were used for participant responses (labeled on the computer keys) because English participants were not trained in phonetic transcription. Participants were trained on the use of these symbols in the first phase of the study. The participants were asked to identify the word from the five choices ("jada," "jjada," "chada," "sada," and "ssada") in a five-alternative forced-choice identification task, thus the target correctness level was 60%. Each of the five utterances was mixed with 200 instances of bubble noise, resulting in 1,000 mixtures. These were divided into five blocks of 200 mixtures for presentation, with breaks provided between blocks. The experiment took each subject approximately 1 hr 30 min. The BPS parameter was controlled globally over all five stimuli.

Results

As an illustrative example, Figure 5 shows importance maps for L1 and L2 Korean speakers on two utterances: the lax affricate /tɕ/ labeled as "j" and the tense affricate /tɕ*/ labeled as "jj." These results suggest that L1 Korean listeners used different cues than the novice English listeners, especially for the lax affricate, /tɕ/. L1 Korean listeners used several cues to identify /tɕɑdɑ/ throughout the duration of the initial consonant and vowel. These cues include VOT at word onset, frication noise above 4 kHz, and formant information between 2 and 3 kHz. In contrast, English novice listeners seemed to focus exclusively on word onset VOT and low-frequency formant transitions between 0.5 and 1 kHz during the initial consonant. For /tɕ*ɑdɑ/, both groups used similar cues, relying on frication noise above 7 kHz, formant transitions, and VOT at word onset. These results appear to agree with our hypothesis that native English speakers would focus on VOT for differentiating tense from lax affricates in Korean. Because VOT is a weak cue for the lax/tense distinction (e.g., C.-W. Kim, 1965), these listeners would be expected to perform poorly in generalizing to a new set of tokens.

Final BPS levels. The average number of bubbles for L2 Korean listeners was 28.7 BPS, and that for native Korean listeners was 25.8 BPS. Thus, both groups showed a similar degree of noise robustness for these tokens. This could be due to the fact that they were required to participate in the pretask training to distinguish the five stimuli to reach 90% accuracy.

Mandarin Tone "Coarticulation"

Mandarin Chinese has four lexical tones, described as high-level (Tone 1), rising (Tone 2), low-dipping (Tone 3), and falling (Tone 4). These are sometimes represented using numbers on a scale from 1 to 5 to characterize the pitch direction over time, with 5 indicating high and 1 indicating low (thus, 55, 35, 214, 51, respectively; Chao, 1968; Howie, 1976; Yang, 2010). These standard notations are shown in Table 2.

The actual realization of the pitch contours, however, is slightly more complicated than indicated by this notation. Many studies indicate that this system is challenging for speakers of nontone languages. For example, Tone 2 has a gentle falling slope before the majority of the syllable rises, which results in a convex shape similar to Tone 3. Tones 2 and 3 are the most challenging pair for nonnative listeners in tone perception tasks (Gottfried & Suiter, 1997; Hao, 2018; Lai & Zhang, 2008; Lee, Tao, & Bond, 2008, 2010; Liu & Samuel, 2004). Shen and Lin (1991) claimed that there is a correlation between the timing and the F0 difference between the onset and the turning point, and listeners identify Tones 2 and 3 based on the interaction of both factors. In identification tasks, errors are mostly due to ambiguities that derive from this correlation. Tones 3 and 4 can also be confused (Gårding, Kratochvil, Svantesson, & Zhang, 1986). Tone 4 has a falling contour, whereas Tone 3 starts with a falling contour, and sometimes is realized without the subsequent rise (Chao, 1968; Duanmu, 2007; Shih, 1988; Yang, 2010).

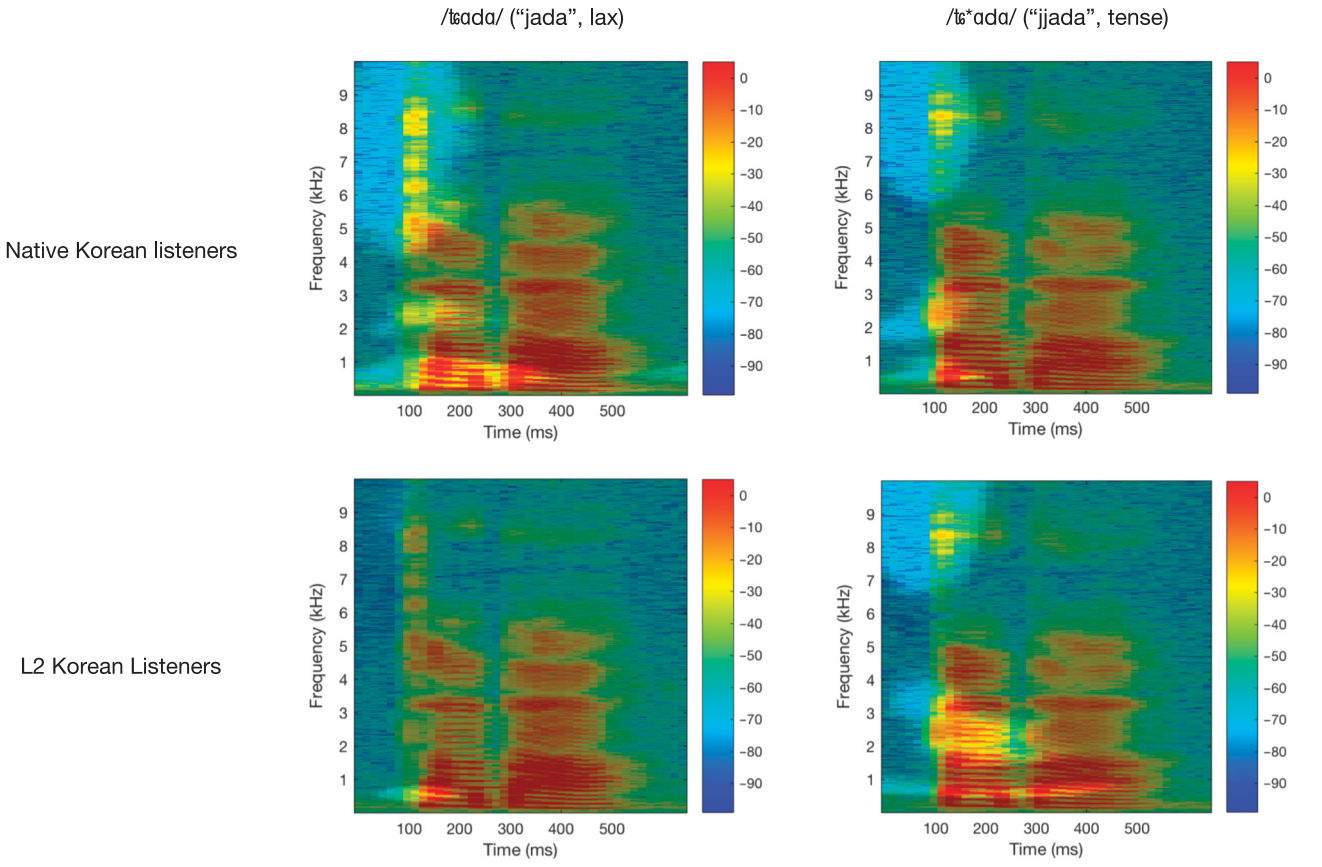
Tone perception is further complicated under coarticulation of surrounding tones (Xu, 1994a, 1994b, 1997), which includes both carryover and anticipatory effects. Carryover effects are the influence of a tone on the following tone and are more salient than anticipatory effects. Carryover effects are mostly assimilatory, resulting in a lower onset tone of a syllable following one with a lower ending tone and vice versa. This deviation is much greater in a "conflicting" context, where the adjacent tonal values disagree (e.g., high final pitch followed by low onset pitch). Native listeners compensate for coarticulation in listening tasks. Huang and Holt (2009) argued that native listeners are highly context dependent in contour tone perception. Jongman and Moore (2000) found that Mandarin listeners normalized for both speaker F0 range and speaking rate when precursors and stimuli varied in the same acoustic dimension, whereas for English listeners, normalization happened as a result of pure acoustic discrimination when both F0 and turning point change.

The current study examined native and nonnative perception of these lexical tone contextual effects using the bubble noise design. We hypothesized that native listeners would be better able to identify the target words than novice listeners. In addition, we expected native listeners to show evidence for making use of the coarticulatory cues in identification and that this evidence would be seen in the importance maps earlier in time (i.e., on the context word form) because of anticipatory effects of coarticulation.

Method

Participants. Five adult native Mandarin speakers and five beginner-level learners of Mandarin whose native language is English or Spanish ("L2 listeners" for short) participated. They ranged in age from 18 to 28 years. All participants provided informed consent. The study was

Figure 5. Native and nonnative importance maps for two Korean affricates. L2 = second language.



conducted under the supervision of the The City University of New York Institutional Review Board.

Stimuli. The stimuli were based on Lee et al. (2008, 2010). Target words were recorded in the context of "tā jiào …" ("his name is …"). Two-syllable names were used in which the first word form was either Tone 1 or Tone 4 and the second word form could be Tone 2, Tone 3, or Tone 4 (see Table 3 for the six possibilities). The first word (syllable) provided the conditioning tone context for the second word, which the subject was asked to identify. All stimuli were considered targets. Multiple tokens of each utterance were recorded in a sound-treated audiometric booth with the built-in microphone of a personal MacBook Pro and then examined to select one token of each target utterance type that closely matched across the six utterances in temporal and intensity patterns. The volume of the stimulus is adjusted by the participants to their comfort level.

Design and procedure. In the three-alternative forced-choice tone identification task, listeners were asked to identify the tone of the last syllable, "wei." The presentation difficulty level began at 15 bubbles per second and was adapted for each token independently based on the participant's performance. Participants listened to 200 trials of each sentence, which was 1,200 trials in total. Because it was a three-way choice, the target accuracy was $\alpha = 66.67\%$.

Table 2. Pinyin notation, five-scale representation of pitch trajectory (5 is the highest and 1 is the lowest; Chao, 1968), and description of the four Mandarin tones.

Tone	Pinyin	Five-scale	Description
1	ā	55	High-level
2	á	35	Rising
3	ǎ	214	Low-dipping
4	à	51	Falling

Results

Figure 6 shows the importance maps of native and L2 listeners' perception of target Tones 2 and 4 in the context of Tone 4 preceding them. The four plots exemplify the differences between participant groups and different target tones. The figure reveals that native listeners made use of pitch information prior to the target word form to facilitate identification of Tone 2. In contrast, identification of the target Tone 4 word used less of the prior context. The L2 listeners' importance maps indicated that tone identity

Table 3. List of sentences used in the Mandarin tone perception experiment showing the tone of the target word "wéi," "wěi," or "wèi" in the context of the tone of the preceding word "zhū" or " zhù." The sentences translate to "His name is zhū wé" for example.

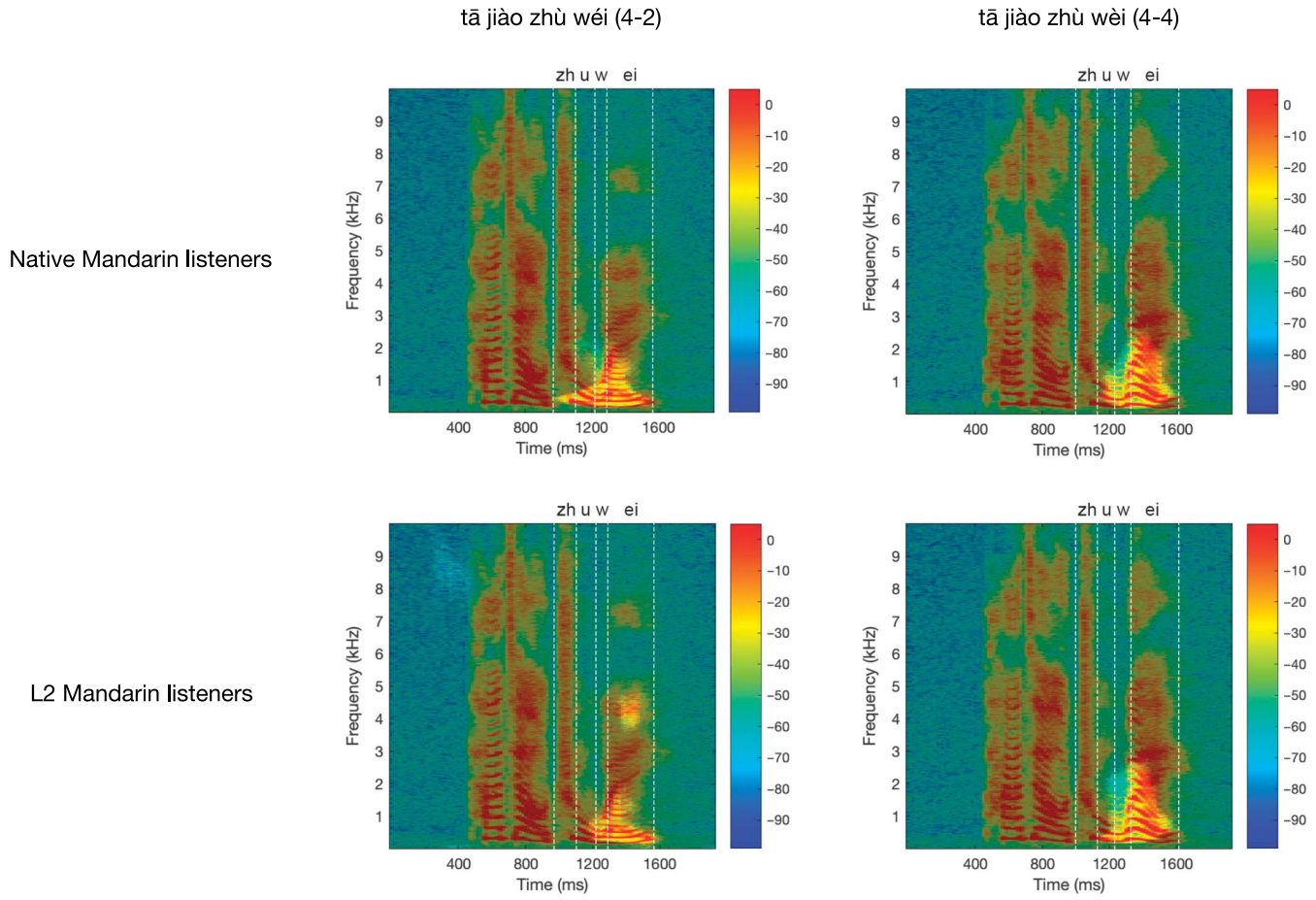
	Preceding context ("zhu")	
Target word ("wei")	**High preceding offset (Tone 1)**	**Low preceding offset (Tone 4)**
Tone 2	tā jiào zhū wéi	tā jiào zhù wéi
Tone 3	tā jiào zhū wěi	tā jiào zhù wěi
Tone 4	tā jiào zhū wèi	tā jiào zhù wèi

was determined from the target word information, and not the prior context.

This pattern of findings supports the literature, which indicates that Tone 2 is often confused with Tone 3, and Tone 2 is considered the most difficult to identify (Gottfried & Suiter, 1997; Lee et al., 2008, 2010). Studies have shown that the prior context influences the tone contour shape (specifically turning point) and the listener's judgments (Jongman & Moore, 2000; Shen & Lin, 1991). The finding that Tone 4 targets did not need information from the prior context is consonant with the claim that Tone 4 is the easiest tone to identify. In addition, English listeners might find Tone 4 relatively easy in final utterance position because it roughly matches the falling pitch contour on declarative sentences in English (Broselow, Hurtig, & Ringen, 1987; Lee et al., 2010; Wang et al., 2006).

Final BPS level. We also conducted a statistical analysis of the final difficulty level. A mixed-design ANOVA was run on the natural logarithm of the final BPS level and showed significant effects in the interaction of context and tone, $F(2, 16) = 12.723$, $p < .01$. Specifically, Tone 3 required a greater amount of the spectrogram to be revealed

Figure 6. Examples of importance maps for native and second-language (L2) listeners for select Mandarin sentences. The maps are similar between the two groups for the 4–4 sentence but show greater use of the preceding context by native listeners for the 4–2 sentence.



than Tone 4 for correct identification. Paired t tests on each tone between the two contexts showed a significant effect on Tone 2 exclusively, $t(9) = -3.038$, $p = .014$, indicating the identification of Tone 2 was highly dependent on its preceding context. Tone 2 following the low offset context (Tone 4) was found to require a greater amount of the spectrogram to be revealed than when it followed a high tone (Tone 1). Additionally, one-way repeated-measures ANOVA was undertaken with tone as the dependent measure in each context condition. Tukey post hoc tests showed that performance was poorest for Tone 3. This is possibly due to the shorter duration of the word when Tone 3 was used (not shown in the figure).

These results illustrate how the bubble noise technique can inform our understanding of which information listeners are using for successful identification. It will be interesting to examine whether training Mandarin L2 learners to focus on the prior context will improve perception of Tone 2 (and Tone 3) in running speech.

Discussion Relative to Other Methods

The question that arises is why use the bubble noise technique rather than one of the traditional designs for studying speech perception? One advantage to this method is that it is highly efficient, in that the important regions of the speech signal of interest can be identified relatively quickly. In addition, this method can provide a useful check of what information listeners are actually using in making decisions in a speech perception task. With native listeners, we often make assumptions about what information is used in successful speech perception. In the case of synthetic speech, in which only the target portion of the signal is manipulated for a study, it can be certain that listeners are using the intended information. However, in the case of naturally varying speech, this may be less certain, as listeners may use secondary cues. The bubble noise technique can reveal whether listeners are doing this, as illustrated in the study with Mandarin listeners.

Of course, other methods continue to be valuable. Editing and/or resynthesizing speech to allow manipulations of circumscribed regions of the signal has played a central role in speech perception research. Over the 70 years of speech perception research (see Liberman et al., 1967; Samuel, 2011), a variety of useful techniques have been implemented to study speech perception, ranging from manipulating the speech information itself (synthesizing or editing acoustic cues), manipulating the context (e.g., presence of visual cues for the "McGurk" effect), overlaying with noise (phonemic restoration), manipulating the lexical status (e.g., real versus nonsense words for the "Ganong" effect), or manipulating the surrounding context (see Samuel, 2011). These methods have been quite successful furthering our understanding of speech perception, but some of them are laborious and require considerable experience. For example, the first studies to demonstrate that AE listeners rely on the initial 40–100 ms of consonant–vowel syllables (assuming a syllable duration of 100–250 ms) to identify place of articulation and that the first- and second-formant transitions are crucial for making this judgment manipulated this information using synthetic speech (Liberman, 1982).

The bubble noise technique is not intended to replace these methods, but to offer another research tool to the toolbox for examining speech perception. This technique follows the spirit of other methods that have overlaid noise on the speech signal to gain insight into how listeners perceive speech. In particular, the bubble noise technique was inspired by the 3D Deep Search method in which speech targets were truncated in time and frequency and masked by varying frequency bands and intensities of white noise (F. Li, Trevino, Menin, & Allen, 2012). In addition, background noise (e.g., speaker babble, shaped noise, white noise) has been frequently used in studies of speech perception in hearing science (Bradlow, 2008; Calandruccio & Doherty, 2007). Adaptive procedures were developed by researchers in the hearing sciences (often to identify hearing thresholds of different types of information; Levitt, 1971; B. C. J. Moore et al., 1986).

What is unique about the bubble noise technique is that it is a particularly efficient method (in terms of time) for identifying portions of the speech signal that are relevant to categorizing information, particularly in exploratory situations where it is less clear (no pun intended!) where in the spectrotemporal space the important information is located. In our view, this method is particularly promising for examining what listeners are doing when they correctly categorize newly learned phonological contrasts. In other words, the method can be used to see if listeners are focusing on secondary cues that covary with the primary cue. For example, AE listeners can use temporal cues to discriminate Japanese long/short vowel pairs, such as /e/ versus /ɛ/ (Hisagi & Strange, 2011). One question that this approach could answer is whether the AE listeners are using the same spectrotemporal regions to successfully discriminate these pairs as native listeners. The method is also likely to be revealing in the study of special populations, such as developmental language disorder and dyslexia. Individuals with these disorders often show poor categorization and discrimination, but it is unclear whether this is due to focusing on the incorrect cues or to higher level issues, perhaps related to attention (e.g., D. R. Moore et al., 2010).

Conclusion

This research note has introduced the bubble noise procedure for measuring "importance maps" of speech in the context of language learning. Through case studies in three language pairs that involve learning new phonetic contrasts, the identified maps show that native and novice listeners rely on different cues even when performing above chance levels. The method shows great promise for tracking how training modifies selective perception in an L2. The method can also be used to examine which information is the most important for perception in more complex contexts,

such as multisyllabic and multiword utterances (Mandel, 2016).

For Hindi learners of English making a /v/–/w/ distinction, the bubble noise method and a more traditional acoustic analysis both found that the F2 onset regions were important. The bubble noise technique, however, is more automated and can be utilized with less training. In Korean, while native and nonnative listeners utilize similar cues to identify the tense affricate /ʧ*/, they use different cues to identify the lax affricate /ʧ/, with L2 listeners focusing only on low-frequency cues, most likely VOT. In Mandarin, while native and nonnative listeners utilize similar cues to identify the fourth tone (in the context of a preceding fourth tone), native listeners utilize more contextual information from a preceding fourth tone in identifying the second tone.

These examples show the utility of the bubble noise method in speech perception research and, in particular, in the study of L2 learning. They suggest that there are differences in the perception of these contrasts between native speakers of a language and learners of that language. Such differences in perception are commonly thought to underlie L2 accentedness in speech production (Flege, 1995; Strange & Shafer, 2008). These differences and our findings with the trained Hindi listener's perception of the /v/–/w/ contrast further suggest that listeners may be able to change their listening "strategies," that is, the cues that they use to recognize these contrasts, over time and with training. The identification of the specific cues that native listeners use suggests a training scheme that might be more effective in changing these strategies, by focusing on those cues that are different between groups and bringing them into alignment. Such approaches will be explored in future work.

There are many clinical implications of using the bubble noise method. The first is that it offers an efficient design for clinicians to examine speech perception in nonnative listeners, providing specific time–frequency regions that are (mis)used. The regions so identified correspond to those identified by traditional acoustic measures. In clinical speech-language pathology, many L2 speakers of English seek services for accent management. To make the accent training effective as well as time efficient, pinpointing specific cues for difficult sound contrasts is critical. Trainers will be able to use information from the bubble noise design and target the specific cues to improve perception and thus production of difficult speech contrasts. For example, in Hindi listeners, based on the acoustic analysis findings and bubble noise findings, training can focus on lip rounding to improve the identification of the specific F2 onset time–frequency region. The second clinical implication is that a comparison of groups based on BPS results allows clinicians to understand the amount of acoustic information nonnative listeners require to identify the target speech sounds accurately. This can also be applied during training as an ongoing measure of progress. The final clinical implication is that clinicians can utilize the bubble noise library in MATLAB to run speech perception experiments of their own design, opening up possibilities for studies that the original authors have not foreseen.

Acknowledgments

This material is based upon work supported by National Science Foundation Grant IIS-1750383 awarded to the first author.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological), 57*(1), 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics, 29*(4), 1165–1188.

Bradlow, A. R. (2008). Training non-native language sound patterns: Lessons from training Japanese adults on the English /r/-/l/ contrast. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 287–308). Amsterdam, the Netherlands: John Benjamins.

Broselow, E., Hurtig, R., & Ringen, C. (1987). The perception of second language prosody. In G. Loup & S. Weinberger (Eds.), *Inter-language phonology: The acquisition of second language sound system*. Cambridge, MA: Newbury House Publishers.

Calandruccio, L., & Doherty, K. A. (2007). Spectral weighting strategies for sentences measured by a correlational method. *The Journal of the Acoustical Society of America, 121*(6), 3827–3836. https://doi.org/10/ft4tkn

Chang, C. B. (2007). Korean fricatives: Production, perception, and laryngeal typology. Retrieved from hhttps://pdfs.semanticscholar.org/2615/7f6d58326753318d945f48d1e6a328e382b8.pdf

Chang, C. B. (2013). The production and perception of coronal fricatives in Seoul Korean: The case for a fourth laryngeal category. *Korean Linguistics, 15*(1), 7–49.

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

Cheon, S. Y., & Anderson, V. B. (2008). Acoustic and perceptual similarities between English and Korean sibilants: Implications for second language acquisition. *Korean Linguistics, 14*(1), 41–64.

Cho, T. (1995). Korean stops and affricates: Acoustic and perceptual characteristics of the following vowel. *The Journal of the Acoustical Society of America, 98*(5), 2891.

Cho, T., & Keating, P. (2001). Articulatory and acoustic studies of domain-initial strengthening in Korean. *Journal of Phonetics, 29*(2), 155–190.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America, 119*(3), 1562. https://doi.org/10.1121/1.2166600

Dart, S. N. (1987). An aerodynamic study of Korean stop consonants: Measurements and modeling. *The Journal of the Acoustical Society of America, 81*(1), 138–147.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology, 55,* 149–179.

Duanmu, S. (2007). *The phonology of standard Chinese*. Oxford, United Kingdom: Oxford University Press.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-linguistic research* (pp. 233–277). Timonium, MD: York Press.

Gårding, E., Kratochvil, P., Svantesson, J.-O., & Zhang, J. (1986). Tone 4 and Tone 3 discrimination in modern standard Chinese. *Language and Speech, 29*(3), 281–293.

GitHub. (2019). Retrieved from https://github.com/mim/auditoryBubbles/

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research, 47*(1–2), 103–138.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research, 41*(17), 2261–2271.

Gottfried, T. L., & Suiter, T. L. (1997). Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones. *Journal of Phonetics, 25*(2), 207–231.

Grover, V. (2016). *Perception and production of /v/ and /w/ in Hindi speakers* (Doctoral dissertation). Retrieved from CUNY Academic Works. https://academicworks.cuny.edu/gc_etds/1545

Grover, V., Shafer, V. L., Campanelli, L., Whalen, D. H., & Levy, E. S. (2019). *Perception of American English consonants /v/ and /w/ by Hindi speakers of English*. Manuscript submitted for publication.

Grover, V., Shafer, V. L., Whalen, D. H., Levy, E., & Kakadelis, S. (2018). Do Hindi speakers of English realize the acoustic cues for American English /v/ and /w/? *The Journal of the Acoustical Society of America, 143,* 1756. https://doi.org/10.1121/1.5035750

Hao, Y.-C. (2018). Second language perception of Mandarin vowels and tones. *Language and Speech, 61*(1), 135–152.

Hisagi, M., & Strange, W. (2011). Perception of Japanese temporally-cued contrasts by American English listeners. *Language and Speech, 54*(2), 241–264. https://doi.org/10/ch72dr

Howie, J. (1976). *Acoustical studies of Mandarin vowels and tones*. Cambridge, United Kingdom: Cambridge University Press.

Huang, J., & Holt, L. L. (2009). General perceptual contributions to lexical tone normalization. *The Journal of the Acoustical Society of America, 125*(6), 3983–3994.

Jongman, A. M., & Moore, C. B. (2000). The role of language experience in speaker and rate normalization processes. In *Proceedings of the International Conference on Spoken Language Processing* (Vol. 1, pp. 62–65).

Kaernbach, C. (1991). Simple adaptive testing with the weighted up-down method. *Perception & Psychophysics, 49*(3), 277–229. https://doi.org/10.3758/BF03214307

Kim, C.-W. (1965). On the autonomy of the intensity feature in stop classification (with special reference to Korean stops). *Word, 21,* 339–359.

Kim, M.-R., Beddor, P. S., & Horrocks, J. (2002). The contribution of consonantal and vocalic information to the perception of Korean initial stops. *Journal of Phonetics, 30*(1), 77–100.

Kim, S. (2000). Post obstruent tensing in Korean: Its status and domain of application. *The Journal of the Acoustical Society of America, 108*(5), 2467.

Lai, Y., & Zhang, J. (2008). Mandarin lexical tone recognition: The gating paradigm. *Kansas Working Papers in Linguistics, 30,* 183–194.

Lee, C.-Y., Tao, L., & Bond, Z. S. (2008). Identification of acoustically modified Mandarin tones by native listeners. *Journal of Phonetics, 36*(4), 537–563.

Lee, C.-Y., Tao, L., & Bond, Z. S. (2010). Identification of acoustically modified Mandarin tones by non-native listeners. *Language and Speech, 53*(2), 217–243.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America, 49*(2B), 467–477.

Li, F., Trevino, A., Menon, A., & Allen, J. B. (2012). A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise. *The Journal of the Acoustical Society of America, 132*(4), 2663–2675.

Liberman, A. M. (1982). On finding that speech is special. *American Psychologist, 37,* 148–167.

Liberman, A. M., Cooper, F., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74,* 431–461.

Liu, S., & Samuel, A. G. (2004). Perception of Mandarin lexical tones when F0 information is neutralized. *Language and Speech, 47*(2), 109–138.

Mandel, M. I., Yoho, S. E., & Healy, E. W. (2016). Measuring time-frequency importance functions of speech with bubble noise. *The Journal of the Acoustical Society of America, 140*(4). https://doi.org/10.1121/1.4964102

Moore, B. C. J., Glasberg, B. R., & Peters, R. W. (1986). Thresholds for hearing mistuned partials as separate tones in harmonic complexes. *The Journal of the Acoustical Society of America, 80*(2), 479–483. https://doi.org/10/b66xpq

Moore, D. R., Ferguson, M. A., Edmondson-Jones, A. M., Ratib, S., & Riley, A. (2010). Nature of auditory processing disorder in children. *Pediatrics, 126*(2), e382–e390.

Park, H. (1999). The phonetic nature of the phonological contrast between the lenis and fortis fricatives in Korean. In *Proceedings of the 14th International Congress of Phonetic Sciences* (pp. 49–72). San Francisco, CA: ICPhS. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_0425.pdf

Samuel, A. G. (2011). Speech perception. *Annual Review of Psychology, 62,* 49–72.

Shen, X. S., & Lin, M. (1991). A perceptual study of Mandarin Tones 2 and 3. *Language and Speech, 34*(2), 145–156.

Shih, C. (1988). Tone and intonation in Mandarin. *Working Papers, Cornell Phonetics Laboratory, 3,* 83–109.

Shin, S. J. (2001). Cross-language speech perception in adults: Discrimination of Korean voiceless stops by English speakers. *Studies in the Linguistic Sciences, 31*(2), 155–166.

Strange, W. (2011). Automatic selective perception (ASP) of first and second language speech: A working model. *Journal of Phonetics, 39*(4), 456–466. https://doi.org/10.1016/j.wocn.2010.09.001

Strange, W., & Shafer, V. L. (2008). Speech perception in second language learners: The re-education of selective perception. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 153–191). Amsterdam, the Netherlands: John Benjamins.

Wang, Y., Jongman, A., & Sereno, J. A. (2006). L2 acquisition and processing of Mandarin tone. In P. Li, L. Tan, E. Bates, & O. Tzeng (Eds.), *Handbook of Chinese psycholinguistics*. Cambridge, United Kingdom: Cambridge University Press.

Xu, Y. (1994a). Asymmetry in contextual tonal variation in Mandarin. *Advances in the study of Chinese language processing, 1,* 383–396.

Xu, Y. (1994b). Production and perception of coarticulated tones. *The Journal of the Acoustical Society of America, 95*(4), 2240–2253.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics, 25,* 61–83.

Yang, B. (2010). *A model of Mandarin tone categories–A study of perception and production* (Doctoral dissertation). Retrieved from https://doi.org/10.17077/etd.su1twqor

Yoon, K. (1999). *Study of Korean alveolar fricatives: An acoustic analysis synthesis and perception experiment*. Paper presented at: Papers of the Mid-America Linguistics Conference, 549–563.