THE ROLE OF LANGUAGE EXPERIENCE IN SPEAKER AND RATE NORMALIZATION PROCESSES

Allard Jongman¹ and Corinne B. Moore²

¹Linguistics Department, University of Kansas, Lawrence, KS 66045, U.S.A.
²Diebold, Inc., 5995 Mayfair Road, North Canton, OH 44720 U.S.A.

ABSTRACT

This study explores the extent to which listeners are sensitive to variations in context when listening to Mandarin tones. Specifically, the effects of speaker F0 and speaking rate are evaluated on the perception of a Tone 2-Tone 3 continuum that varied either along a spectral parameter, a temporal parameter, or both. In addition, two groups of listeners were tested, Chinese and American. Results showed that both listener groups compensate for variations in both F0 and speaking rate. However, Chinese and American listeners did not weight the acoustic cues in the same manner. Results suggest that language background aids in disambiguating phonemic contrasts for Mandarin listeners, but that for English listeners the normalization effects are a consequence of acoustic discriminability. Limitations on perceptual resources allow English listeners to attend to extrinsic information only when intrinsic acoustic differences become more perceptually salient.

1. INTRODUCTION

Early research on vowel perception suggested that context allows the listener to calibrate a vowel space for a given speaker such that subsequent vowel productions can be perceived with reference to that space. Evidence for this notion was first presented in a classic study by Ladefoged and Broadbent [1] and more recently by Johnson [2].

Moore and Jongman [3] recently extended this line of research to the area of tone perception. The study focused on Mandarin Chinese, a language with 4 distinct lexical tones. Mandarin Tones 2 and 3 have very similar contours, which differ primarily in F0 height. It is therefore possible that a Tone 2 produced by a speaker with a low F0 and a Tone 3 produced by a speaker with a high F0 are acoustically very similar. Presumably, then, in order to identify the intended tone correctly, the listener must take into account the speaker that produced the tone.

Moore and Jongman created three synthetic continua between Tones 2 and 3: the first continuum varied only in ΔF0 (the difference in Hz between F0 at onset and valley), a spectral parameter; the second varied only in the position of the ‘Turning Point’ (the point in time where the F0 valley occurs, a temporal parameter); and the third continuum varied in both ΔF0 and Turning Point. Each of these continua was then preceded by a precursor sentence with an F0 range appropriate for either a high F0 speaker or a low F0 speaker. Thus, listeners' responses to a given continuum were compared when it was preceded by a high F0 or by a low F0 sentence.

The continuum that varied only in terms of ΔF0 showed a significant shift in the identification functions: identification of a target tone, ambiguous between Tone 2 and Tone 3, shifted depending on whether the target was preceded by a precursor with a high or low F0 range. Moreover, similar to Ladefoged and Broadbent's results, the effect was contrastive such that the low F0 precursor caused a shift toward the higher Tone 2, while the high F0 precursor caused a shift toward the lower Tone 3.

A similar, significant, contrast effect was observed for the continuum varying in both ΔF0 and Turning Point. Finally, no shift was observed for the continuum that varied only in terms of Turning Point (TP). A high or low F0 precursor did not shift the perception of tone specified by only temporal parameters.

These results suggest that normalization only occurs if context and target vary along the same acoustic dimension. To further investigate this hypothesis, the present study evaluates whether normalization of temporal information also only occurs if context and target both vary along the temporal dimension. The temporal variable that we decided to investigate was speaking rate. Each of the three Tone 2-Tone 3 continua (ΔF0, TP, ΔF0/TP) was preceded by a precursor sentence with either a slow or fast speaking rate.

In addition, this study investigates whether listeners weigh spectral and temporal information differently depending on the linguistic function of this information in their native language. The effects of both F0 and rate variations on tone perception are therefore evaluated by using two different subject populations, speakers of Mandarin Chinese and speakers of American English.

2. EXPERIMENT 1: RATE NORMALIZATION IN TONE PERCEPTION BY NATIVE SPEAKERS
The following three experiments tested the hypothesis that listeners perceive tones in part by normalizing for speaking rate of a precursor. Subjects heard a precursor (fast or slow) plus a stimulus from the Tone 2-Tone 3 continuum described in [3]. The precursors (the segmental context was Zheige zi nian (‘This word is ___’), were modified from a natural utterance of a female speaker. The mean F0 range of the precursor was approximately halfway between the average F0 for the high and low precursors in [3].

Modified versions of the original precursor were created by a sinusoidal method developed by Quatieri and McAulay [4]. The scaling algorithm was applied to the natural utterance to derive an expanded (1.35 x original utterance rate) and a contracted (.65 x original rate) version. Formant frequencies and F0 remained consistent with the original utterance.

The target words (/u/) (Tone 2 ‘not’, Tone 3 ‘dance’) were 11 members of the synthetic Tone 2-Tone 3 continuum varying in ΔF0, Turning Point, or both (as used in [3]), in which ΔF0 varied from 25 Hz to 75 Hz (stepsize: 5 Hz), and Turning Point varied from 20 ms to 220 ms into the tone (stepsize: 20 ms).

A total of 33 native speakers of Mandarin participated in this experiment (ΔF0: 10, TP: 11, ΔF0/TP: 12). The test consisted of a total of 220 trials (11 stimuli x 2 rates x 10 repetitions). Subjects were asked to label each stimulus as either A or B by pressing the button with the appropriate label. Once subjects had heard the pairs again, 10 more pairs were played. All pairs played A first, then B. Subjects who indicated they could hear a difference between A and B went on to task 2, in which 10 repetitions of A and B were presented in isolation at random. A total of 30 monolingual English speakers participated in the experiment consisting of the precursor phrases and the entire continuum of stimuli.

3.1 Results

For the ΔF0/Turning Point continuum there was a significantly earlier shift toward Tone 3 responses in the fast precursor condition (t(11)=2.43, p<.03). For the ΔF0 continuum, no significant difference in responses between the two speaking rates was found (t(10)=1.05, p>.3). Finally, for the Turning Point continuum no significant effect of precursor speaking rate was observed (t(9)=0.86, p>.4).

4. EXPERIMENT 3: SPEAKER NORMALIZATION IN TONE PERCEPTION BY NON-NATIVE LISTENERS

The following experiment examines the hypothesis that speaker normalization is conditioned by language experience by exploring whether speakers of a non-tone language such as English.

A total of 33 native monolingual speakers of English served as subjects (ΔF0: 11, TP: 10, ΔF0/TP: 12). Materials and procedure were virtually identical to those described for Experiment 1. Each experimental session included three tasks. In task 1, subjects listened to the endpoints in pairs 10 times, where the first stimulus in each pair, the Tone 2 endpoint, was labeled "A", and the second stimulus, the Tone 3 endpoint, was labeled "B". Subjects were told to listen to the 10 pairs first. If subjects wanted to hear the pairs again, 10 more pairs were played. All pairs played A first, then B. Subjects who indicated they could hear a difference between A and B went on to task 2, in which 10 repetitions of A and B were presented in isolation at random. Once subjects had reached 80% accuracy on the last 10 stimuli in task 2, they were given the experiment consisting of the precursor phrases and the entire continuum of stimuli.

4.1 Results

For the ΔF0/Turning Point continuum a shift occurred toward low tone responses in the high precursor condition. The difference between responses in the two conditions was
significant \[t(12)= -2.29, p<.04\]. For the \(\Delta F_0\) continuum there was no significant shift in identification as a function of the precursor \[t(7) = -.13, p>.9\]. Finally, for the stimuli varying in Turning Point there was no significant shift as a function of the precursor \[t(8)=1.12, p>.3\].

5. DISCUSSION AND CONCLUSIONS

Included in this study on tone perception were experiments on rate normalization by both Mandarin and English listeners, and on speaker normalization by English listeners. Each experiment tested perception of acoustic cues demonstrated to be relevant to distinguishing Tones 2 and 3 in Mandarin. These acoustic cues were tested independently and in combination. They are: \(\Delta F_0\), a fundamental frequency parameter; Turning Point (TP), a temporal parameter; and both \(\Delta F_0\) and Turning Point together. Table 1 shows the pattern of results for all experiments, including those on speaker normalization by Mandarin listeners reported in [3], according to acoustic cue and normalization type.

There are two major findings of this normalization study, summarized in Table 1. First, Mandarin listeners normalize for both speaker F0 range and speaking rate information in tone perception. The general pattern for the Mandarin listeners is that normalization only occurs when precursors and stimuli vary in the same acoustic dimension. Second, language background influences normalization. The different pattern of findings for the two language groups suggests that language background aids in disambiguating phonemic contrasts for Mandarin listeners, but that for English listeners the normalization effects are a consequence of acoustic discriminability.

<table>
<thead>
<tr>
<th>MANDARIN</th>
<th>(\Delta F_0/TP)</th>
<th>(\Delta F_0)</th>
<th>TP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker normalization</td>
<td>*</td>
<td>**</td>
<td>ns</td>
</tr>
<tr>
<td>Rate normalization</td>
<td>ns</td>
<td>ns</td>
<td>*</td>
</tr>
<tr>
<td>ENGLISH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaker normalization</td>
<td>*</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Rate normalization</td>
<td>*</td>
<td>ns</td>
<td>ns</td>
</tr>
</tbody>
</table>

Table 1: Summary of normalization findings. Precursors were manipulated either in terms of F0 (Speaker normalization) or speaking rate (Rate normalization). Target stimuli varied either in F0 (\(\Delta F_0\)), Turning Point (TP), or both \(\Delta F_0/TP\). Speaker normalization results for Mandarin listeners were reported in [3]. * indicates \(p<.05\), ** indicates \(p<.001\), ns is nonsignificant.

5.1 Speaker and Rate Normalization by Mandarin Listeners

The results of Experiment 1 and those of Moore and Jongman [3] indicate that Mandarin listeners normalize for both F0 range and speaking rate in tone perception. Results also show that Mandarin listeners are sensitive to the degree of acoustic specification of the tone stimuli; the single-cue continuum contained less intrinsic information, elevating the importance of the extrinsic F0 range information. One hypothesis that emerged from the speaker normalization experiment [3] predicted that normalization occurs when both stimuli and precursors vary in the same acoustic dimension. This is because normalization effects were observed for stimuli varying in the F0 dimension (\(\Delta F_0\), and \(\Delta F_0/TP\)), but not for stimuli varying only in the temporal dimension (Turning Point). This hypothesis was tested in Experiment 1 which investigated whether Turning Point is perceived relative to a temporal frame of reference established by speaking rate. Findings confirmed the prediction that Turning Point is perceived with respect to speaking rate, and also supported a corollary prediction, that no rate effect would be observed for the non-temporal, \(\Delta F_0\) continuum. The lack of a rate effect for the \(\Delta F_0/TP\) continuum suggested that the temporal dimension was less informative than the F0 dimension. These results indicate that listeners refer to extrinsic acoustic information only when it is relevant to disambiguating intrinsic cues.
5.2 The Influence of Language Background on Normalization

The second major finding of this study concerns the cross-language results for both speaker and rate normalization, which show that normalization is influenced by language background. The hypothesis of the cross-language experiments was that English listeners, who do not use tones to make lexical distinctions, would show a different pattern of normalization than the Mandarin listeners. This prediction was corroborated by comparing the responses to the single-cue continua. As predicted by the hypothesis, English listeners did not use extrinsic F0 range or rate information in perception of the single-cue continua. The Mandarin listeners, on the other hand, showed the largest effects for those same continua which varied in the same acoustic dimension as the precursors. English listeners showed no such magnitude differences. These results reflect the influence of phonetic categories on normalization for the Mandarin listeners.

Despite differences in results for the single-cue continua, however, both language groups showed normalization effects in perception of the ΔF0/TP continua. Our hypothesis that normalization occurs only when listeners are making phonemic judgments predicted, for example, that only Mandarin listeners would compensate for speaker F0 range. Results showed that English listeners also take extrinsic acoustic information into account, even when identifying non-phonemic contrasts. The English effects in the ΔF0/TP continuum do not constitute phonemic normalization, as it was for the Mandarin listeners. Rather, the English results reflect a different perceptual process from the one employed by the Mandarin listeners, governed by the perceptual salience of the acoustic differences.

A role for perceptual salience is suggested by the finding that English listeners show different patterns of effects for identical acoustic manipulations; that is, no effect in the single-cue continua, but an effect when the manipulations of ΔF0 or TP occur together in the ΔF0/TP continuum. To account for these results, it is proposed that listeners take the extrinsic pitch and timing information into account whenever possible, but that limitations on perceptual resources allow English listeners to attend to extrinsic information only when intrinsic acoustic differences become more perceptually salient. That is, we suggest that for English listeners, the cost of normalizing extrinsic acoustic information was too high in the single-cue continua.

Studies by Mullenix et al. [5] and Sommers et al. [6], as well as Verbrugge et al. [7], demonstrate that perception of speech in multiple-talker conditions results in higher error rates for spoken word recognition and vowel identification. The conclusion drawn from these studies is that adjusting to different talkers and rates is demanding of perceptual resources. If English listeners were attending to the fine acoustic distinctions in the stimuli varying in only one dimension, they may not have had the resources available to attend to the precursor information. When stimuli varied in both the temporal and F0 dimensions, however, listeners had more cues to aid discrimination of the stimuli, and thus may have had perceptual resources left to incorporate the precursor information.

6. REFERENCES


