Speech to song illusion: Evidence from MC
Shuo Zhang
1University of Pittsburgh
shz16@pitt.edu

Abstract
The current project seeks to understand the mechanisms of speech to song illusion from a variety of perspective. In this paper I will (1) offer a re-analysis of the data on speech to song illusion by Deutsch (2008) by giving consideration to the research in speech synthesis; (2) test the possibility (and if possible, mechanism) to create speech-to-song illusions in Mandarin Chinese, by applying the target stability hypothesis (Falk and Rathecke 2010) and Target Approximation Model (Xu 2002); (3) test the Falk-Rathcke (2010) hypothesis on the interval structure in the speech-to-song illusion in Mandarin Chinese.

I Introduction
Deutsch and co-workers (Deutsch 1995, Deutsch et al. 2008) observed that a spoken English sentence can be heard as singing by only repeating the same recorded sentence a few times over, without changing any aspect of the sound signal whatsoever. Deutsch concluded that “the present experiments show that for a phrase to be heard as spoken or as sung, it does not need to have a set of physical properties that are unique to speech, or a different set of physical properties that are unique to song. Rather, we must conclude that, assuming the neural circuitries underlying speech and song are at some point distinct and separate, they can accept the same input, but process the information in different ways so as to produce different outputs”.

In subsequent experiments, Falk & Rathcke (2010) found that this effect is not universal and argued that the physical acoustic properties of the signal do indeed play a role in generating the illusion effect. They proposed that the rhythmic-
future experiment to probe into the robustness of tone pitch intervals in real-time speech (by using corpus methods) and its effect on intelligibility. Furthermore, I demonstrate how the interval structure (1.b) plays a less significant role when the duration condition is satisfied in (1.a) in the perception of speech to song illusion. All speech analysis, synthesis, and manipulation are done on Praat (www.praat.com).

II Analysis of Deutsch’s Data

To analyze Deutsch’s data, I pay particular attention to two data sets. In Deutsch’s experiment, it was interesting to compare the two sets of recordings, one by the group (a) of 11 subjects who were asked to reproduce the phrase after having heard it for ten times, the other by group (b) of 11 subjects who only heard it once but were asked to reproduce the phrase as sung phrase. Their output exhibited striking similarity as is shown in the graph below in average pitch.

Meanwhile, there is striking difference between the output of group (a) and group (c) who only heard the phrase once and were simply asked to repeat:

This difference suggests that by focusing on the acoustic properties (or imitating the prosody) of the speech sentence one might produce the sung phrase (sung phrase is usually produced by exaggerating the acoustic features of the sentence, such as prolonged pitch stability). The following graphs show my calculation of the mean speed of pitch change at a 0.01s interval for the two sentences from above, one from group (a) and the other from group (c).

It is evident from the data in the graph 1 and 2 that the speed of pitch change in sung-like
speech (a) is much lower than those in normal speech (c), thus pointing to the enhanced pitch stability when the sentence is perceived as singing after repeated listening. In other words, there is no surprising that the two groups (a) and (b) should produce similar output because they were covertly instructed (one by the repetition effect and the other by words) to emphasize the acoustic feature of the sentence in their reproduction, and those feature were crucial to the production and perception of singing.

Falk & Rathcke (2010) noticed the link between the speech to song illusion and the research literature on the satiation effect of speech in conditions of repeated listening, and speculated that the repetitive structure of the signal triggers a pattern matching process in the listener that results in a strong impression of the acoustic features of a signal. Indeed, studies of the verbal domain indicated that massive repetition of a speech signal creates a situation in which the linguistic and grammatical meaning lose their importance (the effect known as "semantic or syntactic satiation", [9, 17]). Through the establishment of a recurrent pattern, the loop becomes a rhythmically structured event in which the specific melodic and rhythmical properties of the signal may become more and more salient to the listener until they dominate the perceptual impression (Falk & Rathcke 2010). Here I will give additional evidence from speech synthesis to demonstrate the significance of this effect in speech to song illusion.

Speech perception and speech synthesis research have long argued that there is a noted discrepancy between the physical reality of speech signal in speech production and the perception of speech signal by the listeners. In examining and manipulating the spectrogram of speech sound, research indicated that as complicated as the sound waves are, there were actually a large portion of the information in the physical signal that were not used by the human listeners in the perception of normal connected speech. In other words, human perception only picks up very selected information from speech signal in order to maintain its intelligibility, although it may depend on a number of variables (such as loudness and acoustic environment) just how much information was required to process the information. This has resulted in the boom of the research in speech synthesis known as F0 stylization (Sundberg 1987; 1989; 2006).

For instance, regarding pitch flow, there is a threshold speed of pitch change in pitch perception, and as demonstrated by speech synthesis, only syllables with pitch change under this threshold are assigned a level tone, while those exceed this speed is perceived as having more than one tone movements (without regard to tonal or stress languages). The auditory threshold for pitch variation is known as the glissando threshold $G$. It depends on the amplitude (extent) and the duration of the F0 variation. In hearing experiments using short stimuli, either pure tones or speech-like signals, with repeated presentations, a glissando threshold $G = 0.16/T^2$ was measured (Mertens 2009). In one of the recent advances in F0 stylization, the prosogram by Mertens employs three kinds of transformation and assign most of the syllables in English sentences a level tone.

Although not emphasized, F0 stylization usually takes into account the fact that most speech signals are only heard once in real-life situation: “Some F0 variations are clearly perceived as rises or falls; others go unnoticed unless after repeated listening; still others are simply not perceived at all” (Mertens 2009). “In normal conversation, utterances are heard only once. Given the continuous flow of speech, the listener has no time to reflect on the auditory properties of the signal” (Mertens 2009). In other words, the perception of speech signal after repeated listening is usually not addressed in speech synthesis research in general, and it is thus implied that a different scenario might take place in terms of perception if the signal is repeated. And that is the exact research question of speech to song illusion and research projects like this paper.
In the following section I consider the two-part tonal hypothesis proved to be crucial to the generation of speech to song illusion in the experiments on German by Falk and Rathecke by using preliminary evidence from experiments on generating speech to song illusion in Mandarin Chinese. Here I focus on the first tone and third tone of MC. It is known that in connected speech, tone 1 (high level) and tone 3 (low falling, also known as half-third-tone due to its inconsistency with the citation form in most cases in natural speech), unlike tone 2 (mid rising) and tone 4 (high falling), are more likely to have stable pitches therefore facilitating the generation of the illusion. (or rather because the tone 2 and tone 4 have relatively large intervals in their contours, they are most unlikely to be sung-like in real speech).

### III Target Stability

Recent experiments (Xu 2002; 2004) on the maximum speech of pitch change in speech flow indicated that unlike previously thought, physical limitation does play an important role in speech production and perception. More specifically, the maximum speed of pitch change is actually more frequently reached in natural speech, and human listeners show great robustness to high speed computer synthesized speech while having serious limitations on understanding massive undershoot in human speech. Here I apply results from the Target Approximation model in pitch change. Directly relevant to the current research question on the duration and stability of the tone production, Xu’s Target Approximation model outlined three hypothetical scenarios, all of which commonly occur in any language:

1. There is plenty of time for a target, e.g., when a monophthong vowel is well over 100 ms long, or a static tone is well over 200 ms long—The target value would be attained before the end of its allotted time; and a pseudo steady state at that value might be achieved in the case of static tone or vowel. An illustration of this scenario is shown in Figure 3(a).

2. A target is given just enough time, e.g., when a monophthong vowel is 75 ms long, or a static tone is 150 ms long—Within the allotted time interval, the movement toward the target would proceed continuously, and the target value would not be achieved until the end of its allotted time interval. An illustration of this scenario is shown in Figure 3(b).

3. A target is given insufficient amount of time, e.g., a vowel is much less than 75 ms long, or a static tone is much less than 150 ms long—Within the allotted time interval, again the movement toward the target would proceed continuously; but the target would not be approached even by the end of its allotted time interval. By the time its cycle ends, however, the implementation of the target has to terminate and the implementation of the next target has to start, as is illustrated in Figure 3(c).

![Figure 3](image-url)
rates, fast, mid, and slow, with a broad focus. The pitch intervals of the tones are either 5\textsuperscript{st} or 7\textsuperscript{st} (perfect fourth or perfect fifth), thus facilitating the generation of the effect:

<table>
<thead>
<tr>
<th>tone</th>
<th>mean syllable duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>slow</td>
<td>0.252 0.162 0.267 0.255 0.203 0.273</td>
</tr>
<tr>
<td>mid</td>
<td>0.188 0.072 0.167 0.133 0.132 0.183</td>
</tr>
<tr>
<td>fast</td>
<td>0.072 0.068 0.117 0.112 0.09 0.14</td>
</tr>
</tbody>
</table>

Table 3

Preliminary results of perceptual trial experiments showed that this cutoff value is indeed responsible for the generation of the effect. Only the slow speech rate generated the perceptual shift from speech-like signal to the song-like signal, while the other two do not exhibit this effect no matter how many times the recording is repeated. Moreover, the illusion seems to disappear when synthesized acceleration is applied on the slow rate sentence.

In sum, when the cutoff value of Xu’s Target Approximation model is applied, the target stability is demonstrated to be crucial for the generation of speech to song illusion. This is in agreement with Falk and Rathecke’s work on German. Future perceptual experiments are needed to decide the details of the illusion in MC.

IV Interval Structure

The question of interval structure in the speech to song illusion is proved to be complicated in MC, considering the flexibility of tone production and perception in MC, as well as the interaction between lexical tone and intonation (Shen 1990; Xu 1994). When realizing transitions from one tone to another in real-time speech in MC, for instance, from the third tone to the first tone, because the tones are defined relatively, there is a range of accepted intervals (i.e. all the different intervals that might be interpreted as accepted third tone to first tone transition in real speech). The robustness of this interval is true especially when we consider that MC preserves an over 90% of intelligibility spoken in monotone in a non-noisy background (Xu&Patel&Wang 2010). The change in broad and narrow focus semantically also is reflected in prosody, thus in tone intervals.

For the current study I use the same sample sentence from above. First I ask a question: How is tone interval mapped onto the specific tone relations, such as a transition from tone 3 to tone 1? In a series of trial experiments, the sample sentence (with third tone vs. first tone contrast) was read by male native MC speaker using a wide range of intervals from smaller than 1\textsuperscript{st} to 12\textsuperscript{st}. Results indicate that all intervals produced are accepted as normal speech (presumably heard in different context). One explanation for this is that when third tone is contrasted with the first tone in MC, there is no possible confusion as the other two tones in the inventory are rising and falling tones with large intervals and greater speed of pitch change, with no other level tones to confuse with. Thus, as long as the third tone transits into a first tone in an ascending interval, presumably it can be accepted as intelligible and tolerable in normal speech. (While on the other hand, if a descending interval is heard in this case it inflects a tone accent from a different dialect of MC, while still being intelligible in this case). To confirm this empirically, I intend to build a corpus collecting material in real speech and study the interval range of the level tone contrasts.

More strikingly, in one trial of the above-mentioned reading where the speaker intends to speak in a small interval (such as 1\textsuperscript{st}), we found that in a tone pattern of (3-1-1-1-1-1), the first four words are actually assigned the about same pitch height, while the last two tones are about 1-2\textsuperscript{st} higher (Table 4). In a similar trial, an ascending pattern is found within the realization of the five first tone words, thus assigning the five first tone words different pitch heights in an ascending fashion.
Table 4  Irregular pitch-tone interval mapping in the sample sentence

<table>
<thead>
<tr>
<th>syllable</th>
<th>wo</th>
<th>he</th>
<th>san</th>
<th>bei</th>
<th>ka</th>
<th>fei</th>
</tr>
</thead>
<tbody>
<tr>
<td>central pitch(Hz)</td>
<td>85</td>
<td>85</td>
<td>84</td>
<td>84</td>
<td>88</td>
<td>88</td>
</tr>
<tr>
<td>tone</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The great flexibility of tone intervals in accepted normal speech indicates that the constraint of an interval relation that is similar to the musical scalar relation is little in the generation of speech to song illusion. We may as well speculate that given the sufficient pitch target stability, the occurrence of speech to song illusion depends less on the tone interval structure. Moreover, due to the listener’s ability to adjust perceived pitch in different tuning systems to the ones we’re familiar with (such as interpreting as well-tempered pentatonic scale when hearing a Sundanese Salandro scale by the Western listeners), it is likely that the interval structure does not play a significant role in the illusion in the case of MC. Future perceptual experiments is called for in determining the role of interval structure as well as its complex mapping onto the tone relations.

V Discussion

In conclusion, the current study confirms the Tonal Hypothesis by Falk and Rethecke (2010), and the mechanisms in generating speech to song illusion in Mandarin Chinese. In the meantime, the current study does not negate the conclusion by Deutsch. First of all, we do have to take into account part of the physical properties as a precondition in generating the speech to song illusion; nonetheless, we do acknowledge the fact that the same input of the same physical reality of the signal can generate different output when repeated through the human perceptual system.

There exists a long-standing notion of speech to song continuum by ethnomusicologists, anthropologists, who found in many cultures around the world the existence of the vocal performing genres that lie between speech and song (or in some cases known as heightened speech). In the meantime, as illustrated in this paper, what happens to speech perception when repeated signal is presented to human listeners did not particularly attract scholar attentions until recently. Therefore the research into speech to song illusion may offer us a new passage way to navigate the human audio-perceptual system and brain modularity from speech to song.

VI References

http://bach.arts.kuleuven.be/pmertens/prosogram
MIT Press, 45-55.