Speech-to-song Illusion: Evidence from MC
Appendix

SHUO ZHANG

This appendix is a follow-up discussion/proposal on my presentation at SELLC 2010, "Speech-to-song Illusion: Evidence from MC", a result of my discussion at SELLC and at the Phonetics Lab, Institute of Linguistics, China Academy of Social Sciences, where I am currently a visiting research student. In this section, I further illustrate the implications of studying this illusion on the understanding of the diverging points of music and speech signals, categorical perception, and the possible future studies.

Background Review

In my SELLC 2010 presentation, I first re-analyzed two groups of data in Deutsch's first experiment in English, and revealed that although the illusion appeared to be difficult to explain at first, it is after all not that surprising:

"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)."

I also gave consideration to literature from speech production and perception, as well as speech synthesis. This examination revealed that (1) speech scientists have long noticed the existence of discrepancies between the physical acoustic signal produced by a speaker and the perception of the speech. In other words, a lot of information and detailed contours in the speech signal is simply not perceived or ignored by the listener, who only extract limited acoustic information from the speech well enough to process the meaning for the purpose of communication. This has resulted in the so-called F0 stylization in speech synthesis. (2) Research on speech synthesis usually takes it as a premise (although not stressed) that

"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.

Another relevant line of research comes from Xu(2002, 2006)'s experiments on the limitations of speech production. Xu's experiments revealed that contrary to previous notions, the maximum speed of pitch change in speech production is actually very limited, and that this maximum speed is often reached in real-life connected speech. Thus the Target Approximation model is proposed, and subsequently applied to Mandarin and English.

Speech-to-song illusion as a diverging point in music and linguistic prosody (suprasegmental features)

Although not much attention has been paid to investigating the repeated signal in speech perception (in speech synthesis one time is enough as long as it can be clearly understood), my recent research indicated that studying speech-to-song illusion, i.e. song-like perception of repeated speech-like signal, has profound implications on the diverging points of speech signal and song signal. I will present my evidence in this section.

First of all, Falk & Rectheke (2010)'s research on German and my research in Mandarin Chinese showed that physical acoustic properties of the signal do play a significant role in the perceptual transition from speech to song. Thus, when 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 (Deutsch et al. 2008)".

The first part of the conclusion seemed to contradict my argument. Indeed, F&R (2010) also pointed out in the beginning of their paper that "In contrast to Deutsch et al., we generally assume that acoustic properties do play a role in inducing a perceptual shift from speech to song (the Main Hypothesis). We suggest that the stimulus used for the illusion task has inherently been optimal in its acoustic layout to generate the effect."

I argue here that the physical property of the signal does play a role, while the perception of the signal can be viewed in a different model. Assume there is a speech to song continuum in which the physical properties of the signal vary along the continuum, and there will be a corresponding continuum in the perception level. The continuum on the two levels, however, is different (Figure 1):

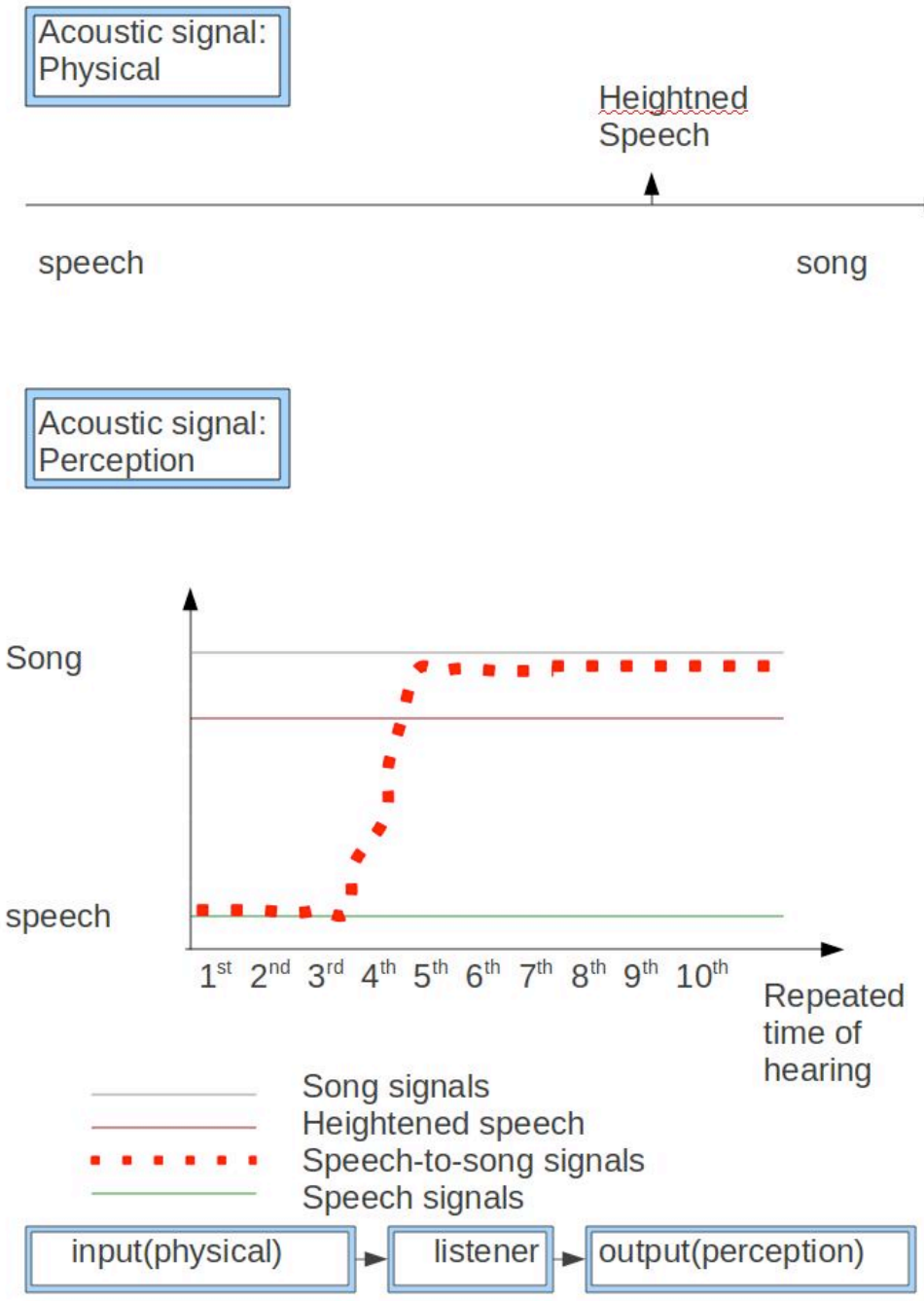

Figure 1 Speech to song continuum: Input-output physical-perception model

According to the previous experiments and my research so far, several physical properties of the acoustic signal may be responsible for generating the effect, including (the list may not be exhaustive):

(1) Target Stability
  a. Target Duration;
  b. Target speed of pitch change;
(2) Target Tonal Make-up (interval relations);
(3) Rhythmic factors;
(4) Pitch Range of the utterance;
(5) Pitch Register;

Using a list of parameters, a table of speech signal, speech-to-song illusion signal, song signal, as well as heightened speech (such as the heightened speech in Peking Opera) can be organized:

Type\Parameter	Target Stability		Interval	Rhythm	Pitch Range	Pitch Register
	Duration	Speed of pitch change				
Song signal	Long	slow	fixed	periodic	large	Low go high
Speech signal(real-life connected speech)	Short	Fast (often reach maximum speed)	random	non-periodic	small	Low-mid(female is higher than male)
Speech-to-song signal	Longer than real-life connected speech(syllble or tone unit over 200ms long)	Significantly slower than max value	Unclear(partly depending on the tone or non-tone languages)	Unclear (did not show effect in F&R 2010)	Unclear(do wide pitch range help generating the illusion?)	Mid to high? (because of the ratio relationship, same interval on the higher register tend to have larger frequency gaps than lower register)
Heigtened Speech (Peking Opera)	Long	Fast	Random	non-periodic	large	Low to high

Table 1 Acoustic Parameters in song, speech, speech-to-song illusion, and heightened speech signals

Table 1 Acoustic Parameters in song, speech, speech-to-song illusion, and heightened speech signals

Therefore, using the physical-to-perception model, speech-to-song illusion offers an opportunity to investigate the physical properties of the diverging points between speech and song signals along the speech to song continuum and their corresponding outputs in perception. Utterances with speech-to-song illusion quality are those with the acoustic properties closer to the diverging points of music and speech prosody, thus the understanding of this phenomenon can shed light on the human categorical perception of well organized acoustic signals such as speech and song.

Through a series of perceptual experiments, the central question to be answered is: "what exactly are these diverging points (in terms of parameters) and how close does it need to be in order to generate the illusion?"

Further implications of this study lies in the understanding of the fundamental differences between human speech prosody and music (specifically singing in this case), or musical melodies (we're comparing the acoustic aspects of speech and song here without reference to other aspects such as syntax and semantics). My hypothesis is that, in a nutshell, given the previously mentioned research on the limitation of speech production and perception, if we set the speech-to-song illusion signals as a central point, song signals tend to prolong (or enlarge) the acoustic features in an more organized fashion (increased target stability, fixed intervals, periodicity in rhythm, widened pitch range and widened pitch register) while on the opposite side, speech signals tend to sacrifice the performance of acoustic parameters (as usually illustrated by the standard or canonical forms such as the textbook four-tone scheme of Mandarin Chinese) in favor of maximum speed of pitch change and effectiveness of communication. This is further proved by the fact that speech allows a greater range of acoustic performance (foreign accent, dialects) as long as utterances can be understood. (Recent research showed that Mandarin Chinese preserves a over 90% of intelligibility even spoken in monotone in a non-noisy background) (Patel & Xu 2010).

Reference

Please see the bibliography section on the original paper for the reference cited.