SPEECH AND SINGING

Shuo Zhang/SPR 11'

I. INTRODUCTION

The research on speech and singing is different than music and language in its scope and subject matter. Speech and singing are most closely connected in that they involve similar articulatory movements (words in the case of speech and lyrics in the case of singing) yet they differ strikingly in terms of prosodic features (i.e. para-linguistic acoustic features such as pitch movements, duration, rhythm, etc). Meanwhile, there are such "grey" areas where the border between speech and singing is vaguely defined. This paper examines speech and singing in three different perspectives: (1) the production of speech and singing; (2) the physical forms of speech and singing (i.e. physical reality of acoustic signal); (3) the perception of speech and singing. Similar and different aspects of the two are compared in these three angles, and the "grey" area along the speech to song continuum is also discussed.

The current study draws the most recent literature from music theory, ethnomusicology, experimental phonetics, cognitive psychology, as well as research on computerized speech synthesis and recognition, and singing synthesis and recognition. In accordance with the research questions of my dissertation, emphasis will be given on the physical characteristics and perception, as well as the discrepancies between the two, whereas singing and speech production is not elaborated in detail.

II. SPEECH AND SINGING: PRODUCTION

Speech and singing production addresses the physiological basis of speaking and singing—essentially the question of how the human biochemical organs work tegother and articulate to produce various kinds of meaningful and organized sounds. It is a topic of discussion for speech sciences and musical sciences, and particularly for speech and hearing departments (which often is related to a science called audiology). It is for this reason that a major part of speech and singing production research is not of particular interest to my field of research (but rather being more relevant to physiology and medical sciences). Here I will briefly outline some relevant research areas from recent literature.

Recent research on speech production in experimental phonetics has paid increasing attention to one factor that was largely overlooked in previous research—the physiological limit of speech production and its role in speech communication (which in turn has profound implications on the diverging points of speech and music perception)[1]. Previous research often tended to assume that these physiological limits are rarely reached and thus provide little value for the understanding of speech production (such as the highest and lowest jaw positions in producing speech). Recently, however, the role of particular constraints has been made clearer, such as in the research on the maximum speed of pitch change in connected speech.

In a series of experiments focusing on Mandarin Chinese, Xu (2002; 2004) tested the ability of human subjects to carry out pitch shifting tasks with the maximum speed of

[1] Although physiological studies are not of my interest in general, this particular study exemplifies as relevant research to speech and song perception as far as I am concerned.

pitch change possible, and revealed that the maximum speed of pitch change is much more often reached in real-life connected speech than previously thought.

Xu's experiments revealed several results that are worth mentioning here: (1)The maximum speed of pitch change varied quite linearly with the size of the pitch change: the larger the size, the faster the maximum speed (equations are derived for the computation of the speed); (2)The minimum time it takes to complete a pitch change is also related to the magnitude of the change, although the correlation is lower than that between the speed and size of pitch change (linear equations are derived accordingly); (3)Compared to the maximum speed of pitch obtained in the study, the speed of pitch change in real speech as reported in previous studies were similar in many cases; (4)in terms of the maximum speed of pitch change, there are no overall differences between native speakers of American English and native speakers of Mandarin Chinese (assuming the latter would carry out more accurately specified pitch shifting between level tones on a daily basis, no difference was observed between the two) (Xu 2004:760-762).

According to these results Xu proposed the Target Approximation Model of pitch change, which was successful in its application to the real-life speech of Mandarin and English as well as to explain certain data from previous experiments such as the co-articulation of Mandarin tones (how the same tone category is realized differently in real-life speech depending on the tone that is preceding the current tone). The model is also applied in my study on speech and song illusion in MC (Zhang 2010), which has profound implications on the perception of speech and song signals. These extended research results from Xu's research will be discussed in the subsequent sections in this paper. Similar research along the line of articulatory constraints and speech rate is seen in

Wu (2010), prosodic phrasing and articulate rate variation (Hansson 2002) and articulatory accounts of speech rhythm in Erickson (2010).

Research on singing production (including those of singing synthesis addressed in the next section) often takes the classical operatic singing as their subject of study (Berndtsson 1995; Bjorkner 2006), although recent years also have seen the increasing interest in non-Western traditions of singing (Wu 2009). It is nonetheless worth pointing out here that in studying the perception of speech and singing, a conceptual division between trained and highly specialized singing forms, and those of everyday casual singing, is necessary (which is why these research on professional singing voices do not directly relate to my research on speech to song illusion). This is similar to the methodological division in speech studies among different speech styles (conversation speech vs. formal speech, for instance, may have very different prosodic features) (See Heldner et al. 2007). Specialized forms of singing differ from everyday casual singing of laypersons in that they involve severe modifications on the timbre of singing voice, obtained through years of training using a specialized technique of singing production. The everyday casual singing, in contrast, usually involves timbre and other characteristics more similar to speech production therefore facilitating our comparison between the two in terms of their diverging points in tonal and temporary dimensions (as proposed in my dissertation project).

Research on singing production often attempts to quantitatively analyze the different singing styles. Stone (2002) examines a female subject with professional experience in both operatic and Broadway styles of singing, who sang examples in these two styles. Further, as a reference point, the style of her speech was also compared. Variation in

styles associated with pitch and vocal loudness was investigated for various parameters: subglottal pressure, closed quotient, glottal leakage, H1 -H2 difference (the level difference between the two lowest partials of the source spectrum), and glottal compliance (the ratio between the air volume displaced in a glottal pulse and the subglottal pressure). Formant frequencies, long-term-average spectrum and vibrato characteristics were also analyzed. Results showed that characteristics of operatic style emerge as distinctly different from Broadway style, the latter being more similar to speaking.

III. SPEECH AND SINGING: PHYSICAL PROPERTIES OF ACOUSTIC SIGNALS

This section focuses on the physical reality of speech and singing sounds: the acoustic properties of the sound signal. I will present the research literature in three complementary parts: analysis by synthesis; analysis on pitch and timbre; analysis on rhythm.

1. Analysis by Synthesis: Rules

Physical description of acoustic signals of speech and singing sounds is acquired through the acoustic analysis using spectrogram softwares. Acoustic analyses, although revealing, tend to supply an overwhelmingly great amount of data (the overwhelming

amount of information presented in acoustic analysis will be specifically discussed in the next section), and the problem is to identify which ones are perceptually relevant. In the light of this researchers have long developed the methodology called "analysis by synthesis", which is, to this day, an irreplaceable tool. The idea of analysis by synthesis comes out of practicality, as suggested by Sundberg (2007):

> "This difficulty of describing sounds is a major problem in music acoustics, since one of its main research areas is the sound of musical instruments. Hence, the ultimate task is to describe and explain how they sound and why they sound as they do. When my interest in music acoustics started, it was common practice in much organology to describe for example, organ timbre in terms of moon shine, or rattling birch leaves… The solution was analysis by synthesis, and I first applied it to the singing voice, the most common of all music instruments. The method implies that you analyze the object by synthesizing it. If you want to describe what characters a singer's voice, you simply synthesize it. As soon as your synthesis contains all the timbral characteristics of the original, you know that from a perceptual point of view your synthesis is exhaustive. If the synthesizer is constructed as an analogue to the vocal apparatus, i.e., if it contains a set of formants attached to a voice source, just as the voice organ, your description is likely to be quite informative…There are several important advantages with the analysis-by-synthesis method. One is that you can find out what acoustic properties are the salient ones. Another advantage is that you do not need a terminology for describing the timbral properties of the instrument. It is enough that you know how the instrument sounds so that you can compare it with the synthesis. A third advantage is that working with sound synthesis tends to draw your attention to details that may be quite important even though mostly unnoticed. Listening to the synthesis helps to direct your attention to such characteristics, and then, it is possible to define a terms for them."

Sundberg (ibid) describes several rules that were derived using this method regarding singing and speaking in terms of their differences. These rules contribute greatly to our understanding of the differences between speech and singing signals:

(1)The rule *duration of consonants* takes into account the fact that in many languages consonant duration depends on length of the vowel preceding it. Thus, a long vowel is followed by short consonants and vice versa. This implies that the duration of a syllable

will be different depending on if it is counted from the onset of the consonant or from the onset of the vowel, as illustrated in Figure 1. Sundberg commented that the principle that tones start at the vowel onset in singing is in accordance with the result of an experiment where non-singer subjects were asked to pronounce syllables in a metrically regular sequence in synchrony with visual and acoustic timing cue appearing at a constant interval (Rapp, 1971). Generally, the subjects synchronized vowel onset with the timing cue, although vowel onset tended to lag behind the time marker in the consonant clusters /str/ and /st/. This delay, as the researchers speculate, may be due to the subjects' lack of training. (ibid:207)

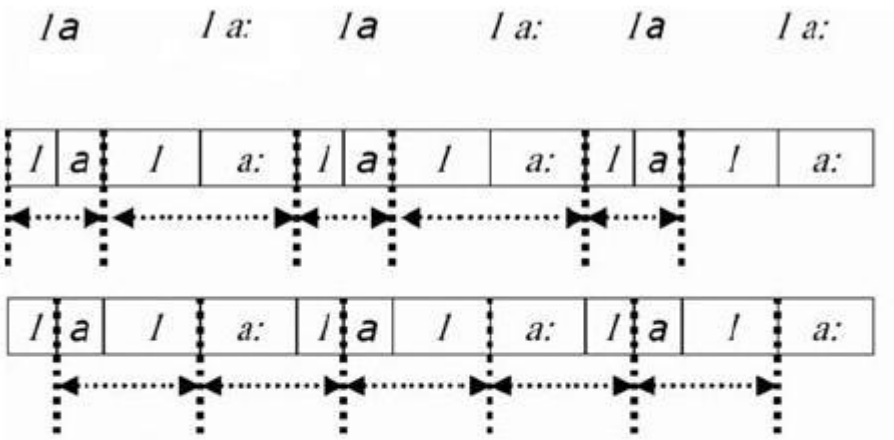

Figure 1

Consequences for timing of definition of syllables. The figure shows the same sequences of the syllable /la/ shown at the top. In the middle row the syllable is defined as a consonant+vowel unit, in the bottom row it is defined as a vowel+consonant unit.

In F&R (2010)'s study on speech to song illusion, this principle is applied as one of the main hypotheses:

> "Segmental grouping: It has been advanced by Sundberg that intervocalic intervals constitute the smallest rhythmic units in sung speech as opposed to syllables in speech. Thus, grouping of intervocalic intervals and not of syllables will lead to an earlier shift from speech to song."

However, the results of F&R's experiments did not support their hypothesis (ibid: 4). In other words, the segmental grouping based on intervocalic intervals did not show an effect of facilitation to the generation of speech to song illusion. My suspicion of this principle lies in its applicability to what I call the "everyday casual singing" (see earlier sections of this paper). Since Sundberg's research was largely based on classical operatic singing (which can be seen from his chosen musical examples to illustrate this very principle-a synthesized phrase from Mozart's opera), it should not be assumed that it can automatically apply to the untrained voice of casual singing but rather should call for an independent study. In addition, different styles of trained singing (such as jazz singing or broadway musical theatre singing) could also yield different results.

(2)*Timing of pitch change* demonstrates a small but important detail of sung performance. The rule states that the pitch change should take place during the consonant preceding the vowel to be sung on the next note. Thus, the new note should begin with its target F0 rather than with an F0 that approaches this target.

(3)The *timbral consequences of a higher larynx* were analyzed by means of a listening test with synthesized ascending scales (Sundberg & Askenfelt, 1983). The results showed that the increase of formant frequencies was the most revealing characteristic of a rising larynx.

(4)*Diphthongs* were realized by letting the formant frequencies of the first vowel remain for 35%of the tone's duration on the values belonging to the first vowel and then starting to approach the formant frequencies of the second vowel in the diphthong.

(5)*Overtone singing* is characterized by the simultaneous appearance of two pitches in the sound produced by one single singer. It seemed reasonable to assume that this was a case of tuning formant frequencies such that they agreed with the frequency of a spectrum partial, which therefore was enhanced. Experiences from the MUSSE synthesizer showed that the effect was rather weak if only the second formant frequency was tuned to the frequency of the partial to be enhanced. Instead the second and third formants were tuned to a cluster with the second one placed on the frequency of the partial. Using this rule, the overtones of a constant drone tone could be made salient.

(6)*Coloratura passages*, i.e., rapid sequences of short notes sung on a single vowel need to be performed in a special way. Analysis of coloratura passages showed that singers make a turn with their F0 around the target F0 values. Thus, each target frequency is represented by a complete vibrato cycle (Sundberg 2007).

These results, derived from analysis by synthesis, are informative and revealing in effectively describing the characters of singing (in some cases in comparison to speech). This method is also widely applied in speech and singing research (Auran et al 2010; Benton 2010). However, I would still like to remain cautious, as I did with the first principle, to not generalize their applicability to other types of singing under the examination of specific studies.

2. Spectrogram Feature Analysis: Pitch and Timbre

Research on pitch and timbral aspects of speech and singing relies on spectrogram analysis of the acoustic signal, from which a quantitative comparison can be derived. Several relevant topics in the recent literature are explored in this section.

The singer's formant (Fs) is a prominent spectrum envelope peak near 3000 Hz that appears in voices sung by trained Western classical singers (which essentially explains why classical opera singers' voice can penetrate the loud sounds of a symphony orchestra and be clearly heard by the audience sitting in the back rolls without using a microphone. The energy concentration for the symphony orchestra usually centers around 500Hz). It is proved that a raising cluster of formant 3, 4, and 5 is especially important since this energy allows singers voices to be heard over the loud orchestra in the big concert or opera (Fant, 1970, Sundberg, 1970). Over the years, various research literatures have investigated the Fs in singing found cross-culturally by using many different methodologies. Wu (2009) investigated the Fs by comparing two completely different training techniques: Trained Chinese opera singing techniques vs. Western classically trained singing techniques. Her conclusion revealed that the singer's formant is indeed present in the case of Peking opera singing, which allows the same kind of effect (singers' voice to be heard over the loud and piercing sounds of *jinghu* and percussion ensembles *luogu*) (ibid: 6).

Research on singer's formant exemplifies the issue of spectrum distribution among different types of vocal genres in similar studies. These results are often used to explain relevant acoustic effects and at times, apply to the acoustic engineering. For instance,

Borch et al.(2002) investigated the frequency and energy distribution of the pop singers in comparison to the loud speaker devices used in performance. It is suggested that pop singers' difficulties to hear their own voices may be reduced if the frequency range 3-4 kHz is boosted in the monitor sound. Cleveland et al (2000) studied the long term average spectrum characteristics of country singers during speaking and singing. The results support the conclusion that the resonance characteristics in speech and singing are similar and that country singing is not characterized by a singer's formant.

Another line of research in this area is to investigate acoustic features (including mainly pitch properties, interval, pitch range, and melody) relating to certain types of pragmatic situations in speech and singing production, such as to distinguish meaning and convey emotional status (either in song or speech), or to signify a certain speech style. Borras-Comes et al (2010) investigated the role of pitch range in establishing intonational contrasts in Catalan. In Catalan, the same rising nuclear pitch accent L+H* is used in three different sentence-types, namely statements, contrastive foci, and echo questions. Using identification tasks and responses (reaction times), this research provides evidence regarding the role of pitch range in relation to pragmatic meaning in speech. Building on existing literature, this represents further evidence that pitch range can be used to make phonological distinctions among a variety of pragmatic meanings, and strengthens the argument that this needs to be represented descriptively at the phonological level. Similar research is also seen in the pitch interval of the knock-knock jokes (Day-O'Connell 2010)[2], the characteristics of infant directed speech (Gustafsson 2002; van de Weiger

[2] Essentially this research proves empirically the hypothesis that knock-knock jokes, albeit the differences among individual performance, are characterized by an interval of approximately a minor third. The author also explained this in terms of physiological constraints of speech production.

2002)[3], the parameters of perceived "good" imitation of foreign language sounds (Persson et al 2007), what constitute features in reading versus spontaneous speech styles (Ramanarayanan et al 2010), and the acoustic properties of what is perceived as a "good speaker" (Strangert 2007).

3. Speech and Singing Rhythm

Speech rhythm has attracted a lot of scholarly interest in recent years due to the new experimental techniques available and the more mature models of viewing speech rhythm without having to rigidly compare speech rhythm with musical rhythm, especially when they have to be regulated by the periodicity (as in music). Three lines of research are identified and will be discussed in this section.

First, rhythmic class and speech perception. Earlier notions of classifying world languages according to their rhythmic characters, particularly the notion of isochrony (which divides world languages into syllable-timed, stress-timed, and mora-timed languages in terms of rhythm) failed empirical tests in the 1980s. Since then, researchers have been searching for new ways to qualify and quantify the perceived rhythmic differences among languages (such as the nPVI, which is proved to be fruitful in both music and language research). A renewed interest, in recent years, has occurred in the area of speech rhythm, however, by arguing that accurate categorization of speech rhythm seems to be a three dimensional problem (durations of vocalic intervals, intervocalic intervals, and speech rate/tempo). Some studies have made provision for

[3] Infant directed speech refers to the speech of adults when they talk to infants, which is observed to be different in many aspects than normal adult-directed speech.

differences in speech tempo by providing metrics with rate normalizing parameters based on the intervocalic intervals or the vocalic intervals (VarcoC & nPVI-V respectively). Focusing on methodological issues, Benton (2010) applies these different metrics on larger corpora of many speakers and more naturally occurring speech. The results show that, while there does seem to be some neutralization of speech rate by VarcoC and nPVI, there may be another way to further normalize without over generalizing. The next steps are to investigate the new metric proposed, the differences between individual speakers in the data, and use more advanced statistics for analysis of normalization. In addition, the notion of rhythmic class, revived by the nPVI studies, continues to draw attention from researchers, with a considerable amount of debates (Arvaniti et al.2010;Botinis 2002).

Second, the music-inspired models of speech rhythm. Although earlier attempts to fit linguistic rhythm into the shoes of musical rhythm was not successful, recent advancement in this area has found new ways to model speech rhythm based on conceptualizations based on multiple musical theoretical foundations. Chow et al (2010) looked at spoken Cantonese with a musical template for phrasal rhythm. Their research takes the premise that although Cantonese lacks stress at the word level, rhythmic patterns are apparent at the sentence level. In order to develop an understanding of this phenomenon, researchers took a sentence and manipulated the syllabic content of several of its target words in order to observe the consequences for rhythmic structure. Overall, it was found that sentence rhythms conformed to simple musical meters. In addition, syllabic durations could become compressed according to small-integer ratios, such as duplets and triplets. Finally, a tendency was observed for sentences to end on a strong beat, a mechanism that was called the Downbeat Rule. Similarly, Brown et al (2010)

proposed a model of speech rhythm inspired by musical conceptions of meter, which he calls "speech is heterometric". The authors posit that *changes in meter* are central to speech rhythm, and thus that speech is "heterometric" rather than isochronous. In addition, two devices are proposed for obviating the need for meter changes within a sentence, both of them involving subdividing component beats: 1) subdivisions according to simple integer ratios, resulting in duplets and triplets; and 2) subdivisions according to complex ratios, resulting in polyrhythms (ibid:9).

Finally, the use of rhythmic properties to indicate pragmatics (emotional status, meaning, speech styles) is also studied in a similar fashion to those discussed in the previous section (as in pitch and timbre). For instance, Engstrand (2002) discussed the relationships of speech rhythm and speech style in casual and elaborated speech. Recordings of careful readings and unscripted monologues in Swedish and Argentinean Spanish were acoustically analyzed to test the hypothesis that temporal equalization of syllable-sized contoid-vocoid (CV) sequences typically occurs in casual as opposed to elaborated speaking styles. Results support the view that rhythmic patterns associated with stress and syllable-timing may arise as consequences of more primary phonetic intentions rather than themselves representing such intentions.

IV. SPEECH AND SINGING: PERCEPTION

Cognitive sciences take the basic premise that everything that we perceive is processed in relation to the things that we already perceived previously. In the study of speech and singing perception, this principle also constitutes the fundamental research

question, that is, how is our perception of sound processed in relation to the sound patterns that we are already familiar with (often through years of ear-tuning)? Because music and language are the most intensive systematic sound that human beings grow up with, and both involve highly selective inventories of sound elements, how our musical and linguistic background influences/fine tunes our perception of speech sound, singing sound, and non-speech sound becomes essential. In this section I outline two areas of research that are of significance to answering this question.

Perceptual grouping has traditionally been thought to be governed by innate, universal principles.[4] However, recent work has found differences in Japanese and English speakers' non-linguistic perceptual grouping, implicating language influence in non-linguistic perceptual processes (Iversen, Patel, & Ohgushi, 2008). Yoshida et al (2010) proposed two experiments to test Japanese and English-learning infants of 5–6 and 7–8 months of age to explore the development of grouping preferences. The results reveal an early difference in non-linguistic perception between infants growing up in different language environments, quite similar to those found in studies with adults. The possibility that infants' linguistic phrasal grouping is bootstrapped by abstract perceptual principles is discussed.

In another experiment, Bent et al (2006) tested Mandarin and English listeners on a range of auditory tasks to investigate whether long-term linguistic experience influences the cognitive processing of non-speech sounds. As expected, Mandarin listeners identified Mandarin tones significantly more accurately than English listeners; however, performance did not differ across the listener groups on a pitch discrimination task

[4] Perceptual grouping refers to the task where one hears a repeated sound sequence of -short-long-short-long- and is asked to mark the boundaries of the "groups" (i.e., short-long groups or long-short groups). Perceptual bias is thought to be due to patterns subject to the linguistic and musical sound pattern influence.

requiring fine-grained discrimination of simple non-speech sounds. The crucial finding was that cross-language differences emerged on a non-speech pitch contour identification task: The Mandarin listeners more often misidentified flat and falling pitch contours than the English listeners in a manner that could be related to specific features of the sound structure of Mandarin, which suggests that the effect of linguistic experience extends to nonspeech processing under certain stimulus and task conditions.

V. SPEECH AND SINGING: ACOUSTIC PROPERTIES VS. PERCEPTION

In this section I focus on the recent advancement in the understanding of the complex relationships between speech production and perception. Particularly, I focus on the recent understanding of how the acoustic signals produced by human speech are perceived and interpreted differently by the human listeners, a process where a large portion of the acoustic information is lost but nevertheless does not hinder effective communication. Implications on singing are also discussed.

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 (Xu 2002; Xu 2004; Mertens2009; Sundberg 2007). In examining and manipulating the spectrogram of speech sound, research indicated that as complicated as the sound waves are in their raw format, 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 (Xu 2002). 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.[5] This has resulted in the boom of the research in speech synthesis known as F0 stylization (Sundberg 1987; 1989; 2006), a process where the F0 contour of speech signal is synthesized using simpler contour patterns and algorithms while maintaining high intelligibility.

For instance, regarding pitch flow, it was advanced by Mertens (2009) that 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 works 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

[5] It is also a widely acknowledged fact that in the computer synthesized speech only very few formants are used while most of the harmonics are left out, while the speech signal still remains highly intelligible. In certain cases (such as telephone signal filtering), even the Fo (fundamental frequency) is omitted.

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" (ibid: 2). 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—which is the exact theoretical foundation for the research on speech-to-song illusion (Zhang 2010).

In my previous study (Zhang 2010), I showed the significance of this process in the understanding of a perceptual illusion, the speech-to-song illusion, first discovered by Diana Deutsch (2007). In this illusion, recording of a spoken English sentence is played back repeatedly for several times before it is transformed into sung by the listeners' perception. 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-metrical and tonal features of the sample sentence might be crucial factors in such processes. Their work on speech to song illusion in German showed that (1) despite individual differences, the

speech-to-song illusion is a robust perceptual phenomenon comparable to those known in visual perception; (2) acoustic parameters – especially tonal structure – facilitate the perceptual shift from speech to song pointing to an acoustically guided decoding strategy for speech- vs. song-like signals. In my previous work (Zhang 2010), I gave consideration to literature from speech production and perception, as well as speech synthesis to explain the perceptual basis for this illusion. I also discussed the possibility to generate the illusion in Mandarin Chinese, a tone language. A physical-to-perception model is proposed to study this phenomenon, in which I argue that while the physical property of the signal does play a role, 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 continuums on the two levels, however, are different:

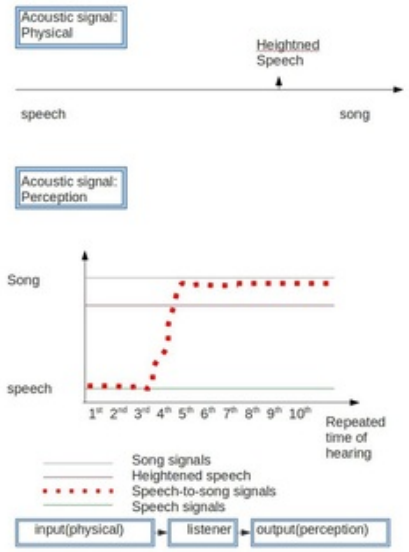

Figure 2. The Physical to Perception Model

Based on this model, I proposed several physical properties of the acoustic signals that 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 Makeup

(interval relations);

(3)Rhythmic factors

(4)Pitch Range of the utterance

(5)Pitch Register

Therefore, using the physical to perception model, I propose that 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 speechtosong 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.

VI. CONCLUDING REMARKS

Research on speech and singing has offered much insights in an attempt to understand these two closely related yet distinctive modes of vocal production. Further implications of this endeavor, as my current research is attempting to achieve, lies in the understanding of the fundamental differences between human speech prosody and music

(specifically singing in this case), or sung 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) (Xu et al 2010).

Preface to Speech and Singing Bibliogrpahy

The research literature on speech and singing entails a highly interdisciplinary endeavor, with innovative studies coming from music theory, (ethno) musicology, music cognition, speech and singing synthesis/recognition, brain sciences, speech and hearing sciences, biological sciences, cognitive psychology, computer science, to name a few. Thanks to this magnitude of interests and diverse perspectives, recent years have seen a increasing amount of speech research that connect speech prosody more closely to the study of singing/musical "prosody" (which essentially is what makes singing what it is other than speech). It is for this reason that our understanding of speech prosody—the traditionally ignored "para"-linguistic components of speech have improved greatly and in turn facilitating better understanding of the singing in relation to speech.

This bibliography has taken a large amount of its entries from several major sources specializing in relevant areas, notably, the Speech Prosody Chicago Conference 2010 (which is the single biggest conference in the world devoted to the speech prosody and relevant research), and the *KTH Speech, Music and Hearing Quarterly Progress and Status Report*(KTH, the national research institution in Sweden, has been the leading institution in the world in developing and research in speech technology and singing synthesis). These two sources represent the most authoritative and up-to-date research on speech and singing. Classical research literature in the area (often coming from KTH as well, represented by research by Sundberg and colleagues) from the 1980s and 1990s are also included as the theoretical foundation for discussion.

Other entries are selected according to the relevant areas of discussion from major journals in the fields.

A key point that needs to be addressed here is the gaps currently existing in the field of speech and singing study—the gap between our current understanding of the two and a more integrated picture of speech and singing, namely, the need for research projects based on non-classical and non-Western forms of music. As I have already demonstrated in this paper, the investigations into these unexplored areas are crucial in order to get a more thorough understanding of many principles underlying human speech and singing, especially if we want to generalize these principles.

Annotated Bibliography: Speech and Singing

***Arvaniti, A. and Ross, T. Rhythm classes and speech perception.Paper presented at Speech Prosody Conference, Chicago, May 2010.**

This study indirectly tests whether American, Greek and Korean listeners can classify low-pass filtered utterances of English, German, Greek, Italian, Korean and Spanish into rhythm classes, by examining how they rate each utterance's rhythm in comparison to a series of non-speech trochees. Such classification was difficult for all groups of listeners and did not support the rhythmic classification of the languages of the stimuli, casting doubt on the impressionistic basis of the rhythm class hypothesis.

***Auran, C. and Bouzon, C. A multi-level approach to speech rate in British English:towards an analysis-by-synthesis method.Paper presented at Speech Prosody Conference, Chicago, May 2010.**

This paper provides a detailed account of the durational differences induced at different structural levels (inter-silence segments, rhythmic units, syllables, syllabic constituents, phones) by changes in speech rate from normal to slow speech in read British English. Using the data described in this study, preliminary results were presented concerning a regression tree model predicting phone durations in slow speech with an average precision of 16 ms in slow speech.

Benton, M. A. Preliminary Analysis of the Relationship of Speech Rate to Speech-Timing Metrics as applied to Large Corpora of Non-Laboratory Speech in English and Chinese Broadcast News. Paper presented at Speech Prosody Conference, Chicago, May 2010.

A renewed interest, in recent years, has occurred in the area of speech rhythm (traditionally defined by categories of speech timing patterns based on perceptual or

acoustic durations of stresses, syllables, or moras). Since accurate categorization seems to be a three dimensional problem (durations of vocalic intervals, intervocalic intervals, and speech rate/tempo), some studies have made provision for differences in speech tempo by providing metrics with rate normalizing parameters based on the intervocalic intervals or the vocalic intervals (VarcoC & nPVI-V respectively). This study applies these different metrics on larger corpora of many speakers and more naturally occurring speech.

***Berndtsson, G. The KTH rule system for singing synthesis. STL-QPSR, (11),1995.**

This article contains a description of rules controlling the singing synthesis at the Department of Speech Communication and Music Acoustics at KTH.The rules controlling the singing synthesizer MUSSE DIG are implemented in Va programming environment originally developed for a text-to-speech system. There are context dependent rules for pronunciation of vowels and consonants, as well as rules for musical performance.In this article, musical rules, general rules for consonants and vowels, and for some special singing techniques are described.

***Bjorkner,E. Why so different? Aspects of voice characteristics in operatic and musical theatre singing. Ph.D thesis,KTH, 2006.**

This thesis addresses aspects of voice characteristics in operatic and musical theatre singing. The common aim of the studies was to identify respiratory, phonatory and resonatory characteristics accounting for salient voice timbre differences between singing styles. Singing voices are also compared with speech voice timbre.

Borch,Z, and D., & Sundberg, J. (2002). Spectral distribution of solo voice and accompaniment in pop music. TMH-QPSR, 43(1), 031-035.

The frequency and energy distribution of the singers is investigated in comparison to the loud speaker devices used in performance. It is suggested that pop singers' difficulties to hear their own voices may be reduced if the
frequency range 3-4 kHz is boosted in the monitor sound.

Borràs-Comes, J; del Mar Vanrell, M; Prieto, P.The role of pitch range in establishing intonational contrasts in Catalan.Paper presented at Speech Prosody Conference, Chicago, May 2010.

In Catalan, the same rising nuclear pitch accent L+H* is used in three different sentence-types, namely statements, contrastive foci, and echo questions. Using identification tasks and responses (reaction times), this research provides evidence regarding the role of pitch range in relation to pragmatic meaning in speech. This represents further evidence that pitch range can be used to make phonological distinctions between a variety of pragmatic meanings, and strengthens the argument that this needs to be represented descriptively at the phonological level.

Botinis, A. (2002). Prosodic effects and crosslinguistic segmental durations. Proceedings of Fonetik, TMH-QPSR, 44(1), 077-080.

The present study is an experimental investigation of the effects of syllable position, stress, focus and tempo on segmental durations in American English, British English, Greek and Swedish. Nonsense disyllabic CVCV words were produced in a carrier sentence under different conditions of stress, focus and tempo. The results indicate that stress and tempo have a major effect on both consonant and vowel across all four languages, whereas the effects of syllable position and focus are hardly evident. Significant interactions were mostly found
between syllable position and stress for the vowel.

***Brown,S.; Weishaar.K.Speech is "heterometric": The changing rhythms of speech.Paper presented at Speech Prosody Conference, Chicago, May 2010.**

This paper presents a model of speech rhythm inspired by musical conceptions of meter. The authors posit that changes in meter are central to speech rhythm, and thus that speech is "heterometric" rather than isochronous. In addition, two devices are proposed for obviating the need for meter changes within a sentence, both of them involving subdividing component beats: 1) subdivisions according to simple integer ratios, resulting in duplets and triplets; and 2) subdivisions according to complex ratios, resulting in polyrhythms.

***Chow, I; Brown, S; Poon, M; and Weishaar, K. A Musical Template for Phrasal Rhythm in Spoken Cantonese.Paper presented at Speech Prosody Conference, Chicago, May 2010.**

This paper takes the premise that although Cantonese lacks stress at the word level, rhythmic patterns are apparent at the sentence level. In order to develop an understanding of this phenomenon, researchers took a sentence and manipulated the syllabic content of several of its target words in order to observe the consequences for rhythmic structure. Overall, it was found that sentence rhythms conformed to simple musical meters. In addition, syllabic durations could become compressed according to small-integer ratios, such as duplets and triplets. Finally,a tendency was observed for sentences to end on a strong beat, a mechanism that was called the Downbeat Rule.

Cleveland, T. F., Sundberg, J., & Stone, R. E. (2000). Long-term-average spectrum characteristics of country singers during speaking and singing. TMH-QPSR, 41(2-3), 089-094.

This paper investigates the acoustic properties of country singers' singing and speaking. The spectral characteristics of country singers' speech and singing were similar. The results support the conclusion that the resonance characteristics in speech and singing are

similar and that country singing is not characterized by a singer's formant.

***Cooper, A and Wang, Y. The role of musical experience in Cantonese lexical tone perception by native speakers of Thai. Paper presented at Speech Prosody Conference, Chicago, May 2010.**

Adult non-native perception is subject to influence from a variety of factors, including native language and musical experience. The present study investigates the influence of these two factors in the perception and learning of non-native lexical tones. Native Thai-speaking musicians and non- musicians completed pre- and post-test identification tasks on five Cantonese tones, with 4 days of lexical identification training. Higher identification accuracy scores for musicians suggest that extensive experience with musical pitch enhances perception of non-native linguistic pitch. However, patterns of tonal accuracy improvement were similar across groups and can be attributed to the influence of the L1 tonal system.

Day-O'Connell, J. "Minor Third, Who?": The Intonation of the Knock-Knock Joke.Paper presented at Speech Prosody Conference, Chicago, May 2010.

In an effort to examine the intonational phenomenon of stylized intonation, knock-knock jokes were collected and phonetically analyzed. Results showed intonation that varied considerably from subject to subject but which was nevertheless constrained in a way that supports a hitherto unexamined supposition among musicologists and linguists: that stylized intonation is defined by the musical interval of a minor third. Results showed a preference for intervals approximating a minor third, as well as an unexpected "boundary" role for the minor third itself, which is interpreted as a consequence of physiology.

Engstrand, O. (2002). Duration of syllable-sized units in casual and elaborated speech: Cross-language observations on Swedish and Spanish. Proceedings of Fonetik, TMH-QPSR, 44(1), 069-072.

This paper discusses the relationships of speech rhythm and speech style. Recordings of careful readings and unscripted monologues in Swedish and Argentinean Spanish were auditorily and acoustically analyzed to test the hypothesis that temporal equalization of syllable-sized contoid-vocoid (CV) sequences typically occurs in casual as opposed to elaborated speaking styles. Results support the view that rhythmic patterns associated with stress- and syllable-timing may arise as consequences of more primary phonetic intentions rather than themselves representing such intentions.

Erickson, D. An articulatory account of rhythm, prominence, and phrasal organization.Paper presented at Speech Prosody Conference, Chicago, May 2010.

This paper examines some articulatory and acoustic characteristics of American English. The results suggest that the jaw may be the articulatory organizer of phrasal rhythm, manifested acoustically through the F2-F1 pattern. Utterance prominence, such as contrastive emphasis, is additionally manifested by increased F0 along with increased duration on the prominent word. The rhythmical organization of the utterance, based on strong-weak jaw opening patterns, may be different from the intonational organization involving pitch accents/ boundary strengths. The author argues that American English prosody might be best described using a parallel system involving both a rhythm system based on articulation, and an intonational system involving pitch notations.

***Falk, S. and Rathcke, T. (2010). On the Speech-To-Song Illusion: Evidence from German. Paper presented at Speech Prosody Conference, Chicago, May 2010.**

The present study investigates the boundaries of speech and song from an acoustic-perceptual perspective. Two hypotheses regarding the tonal and rhythmic-metrical properties of the signal are tested in German.

Fant, G., & Kruckenberg, A. (2006). Individual and contextual variations of prosodic parameters. TMH-QPSR, 48(1), 005-009.

Two aspects of co-variation are investigated. One is inherent in the speech production mechanism, in particular the voice source and its dependency of subglottal pressure and F0. These relations explain how speech intensity co-varies with F0 in connected speech, which has led us to define a mid- frequency F0r in a speaker's available intonation range.

Forsell, M., Elenius, K., & Laukka, P. (2007). Acoustic correlates of frustration in spontaneous speech. Proceedings of Fonetik, TMH-QPSR, 50(1), 37-40.

This paper investigates the acoustic attributes of frustration in spontaneous speech. The speech material was recorded from real life Swedish telephone services by the company Voice Provider. Significant differences between the acoustic cues for neutral and emotional speech were discussed. Anger was characterized by a rise of fundamental frequency and an increase in speech amplitude, whereas despondency reduced the syllable rate significantly. The emotional intensity raised the pitch, increased the amplitude and decreased the syllable rate. Correlations were also found between perceived emotions and acoustic speech parameters.

***Globerson, E; Lavidor, M. Golan, O; Kishon-Rabin, L; Amir, N. Psychoacoustic abilities as predictors of vocal emotion recognition. Paper presented at Speech Prosody Conference, Chicago, May 2010.**

The mechanisms underlying vocal emotion recognition (VER) have been the subject of extensive research in the last decades. Evidence supporting a linkage between the level of musical background and vocal emotion recognition abilities was indicated in several studies, while others pointed to a linkage between Theory of Mind/emotional intelligence and VER. In the current paper we highlight pitch discrimination abilities as successful

predictors of VER.

***Gustafsson, L. (2002). Assessing F0 patterns in infant-directed speech: A tentative stochastic model. Proceedings of Fonetik, TMH-QPSR, 44(1), 061-063.**

This paper reports an attempt to detect possible regularities in the patterns of F0- contours typically occurring in infant-directed speech (IDS). The study used a limited sample of IDS to attempt to disclose possible recurrent patterns of F0 variation regardless of their underlying linguistic and phonetic significance. The variation of the original F0 contours was quantified in steps representing different ranges of F0 slopes.

***Halle, P; Chang, Y.C; and Best, C. Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners, Journal of Phonetics, in press.**

This paper is a cross-linguistic study comparing Taiwan Mandarin and French listeners in terms of the categorical nature of perception of tone contrasts by native listeners of tone languages. Listeners are tested on three tone continua derived from natural Mandarin utterances within carrier sentences, created via a pitch-scaling technique in which within-continuum interpolation was applied to both f 0 and intensity contours. Classic assessments of categorization and discrimination of each tone continuum were conducted with both groups of listeners.Results indicated that Taiwanese listeners' perception of tones is quasi-categorical whereas French listeners' is psychophysically based.The findings suggest that despite the lack of lexical tone contrasts in the French language, French listeners are not absolutely ''deaf'' to tonal variations. They simply fail to perceive tones along the lines of a well-defined and finite set of linguistic categories.

***Hansson, P. (2002). Prosodic phrasing and articulation rate variation. Proceedings of Fonetik, TMH-QPSR, 44(1), 173-176.**

In this paper, results from two studies on articulation rate variation and prosodic phrasing are presented. Production data are presented that suggest a progressive articulation rate reduction over the course of the prosodic phrase in southern Swedish, and results from a perception experiment reveal the importance of articulation rate reduction for perceived boundary strength.

***Heldner, M., & Edlund, J. (2007). What turns speech into conversation? A project description. Proceedings of Fonetik, TMH-QPSR, 50(1), 45-48.**

The project Vad gör tal till samtal? (What turns speech into conversation?) takes as its starting point that while conversation must be considered the primary kind of speech, we are still far better at modelling monologue than dialogue, in theory as well as for speech technology applications. Through this project, features that are specific to human-human conversation – features that turns speech into conversation are investigated. Acoustic and prosodic aspects of such features are discussed.

Karnebäck, S. (2004). Speech/Music Discrimination Using Discrete Hidden Markov Models. TMH-QPSR, 46(1), 041-059.

A speech/music discrimination system using discrete Hidden Markov Models has been designed. The system has been evaluated using separate training, development and test databases. Different features in speech and music signals .were discussed.

***Lamarche, A. Putting the Singing Voice on the Map:Towards Improving the Quantitative Evaluation of Voice Status in Professional Female Singers. Ph.D Thesis, KTH,2009.**

Diagnostic and evaluative methods used in voice care are mostly designed for the speaking voice, and are not necessarily directly applicable to the singing voice. This thesis investigated the possibilities of fine tuning, im- proving and quantifying the voice status assessment of the singer, focusing especially on the Western operatic female voice.Detailed discussions on the aspects of features of the singing voice are presented.

Martin, P. Prosodic structure revisited: a cognitive approach-The example of French. Paper presented at Speech Prosody Conference, Chicago, May 2010.

This paper looks in some details into the mechanism of the decoding process between speaker and listener in normal conversations, focusing on the role of prosodic events in the specific case of French. Possible comparisons with music can be derived in conjunction with other research.

Meireles, A; Tozettil, J; and Borges, R. Speech rate and rhythmic variation in Brazilian Portuguese. Paper presented at Speech Prosody Conference, Chicago, May 2010.

This paper discusses new inventive methodologies for the classification of speech rhythms. Moreover, it sheds new light on the search for isochrony in speech by showing that fast rates exacerbates the timing characteristics of rhythms, i.e., syllable- and/or stress-timing properties are more easily perceived at these rates.

***Mürbe, D., Pabst, F., Hofmann, G., & Sundberg, J. (2002). Effects of a professional solo singer education on auditory and kinesthetic feedback - a longitudinal study of singers pitch control. TMH-QPSR, 43(1), 081-087.**

The significance of auditory and kinesthetic feedback to pitch control in singing was described in a previous report of this project for students at the beginning of their

professional solo singer education (Mürbe et al., 2002). Since it seems reasonable to assume that pitch control can be improved by training, the same students were reinvestigated after 3 years of professional training. The results support the assumption that the kinesthetic feedback contributes substantially to intonation accuracy.

***Patel, A., 2008. Music, language and the brain. Oxford: University Press.**

The most comprehensive book on music and language (with chapters on speech and singing production and perception) to date.

***Peretz, I., in press. Music, language and modularity in action. In Rebuschat, P., Rohrmeier, M., Hawkins, J. and Cross, I. [Eds], Language and Music as cognitive systems. Oxford: University Press.**

In this paper, Peretz examines to what extend music and speech share processing components by focusing on vocal production, that is, singing and speaking, and the role of brain modularity. Relevant literatures are critically reviewed and a tentative conclusion is discussed.

Persson, J., & Westholm, L. (2007). The Parrot Effect – a study of the ability to imitate a foreign language. Proceedings of Fonetik, TMH-QPSR, 50(1).

This experiment attempts to answer the question: how good are people at imitating a foreign language? The results show that it is not completely necessary to get all the phonetic segments of the word right as long as you match the duration of the word and the manner in which the word was originally spoken. The parameters of speech are analyzed in relation to the performance of imitation.

***Pilotti, M., Antrobus, J.S. and Duff, M. (1997). The effect of presemantic acoustic adaptation on semantic 'satiation'. Memory and Cognition 25 (3), 305-312.**

A decrement in the strength of the meaning of a word after rapid repetition of that word has been called “semantic satiation.” This study asked whether this “satiation” might be produced by pre-semantic acoustic adaptation. Category words were utilized to prime the meaning of target words. The adaptation or “satiation” procedure, 30 rapid repetitions of the primes, was compared with a control condition of 3 repetitions.This study concluded that the semantic “satiation” observed here was a decrement in the activation level of semantic representations induced by presemantic acoustic adaptation.

Ramanarayanan, V; Byrd, D; Goldstein, L;and Narayanan, S.Joint Acoustic-Articulatory Study Of Nasal Spectral Reduction in Read Versus Spontaneous Speaking Styles.Paper presented at Speech Prosody Conference, Chicago, May 2010.

Speech styles are one of the primary phenomena of prosodic variation in speech. This paper presents a novel automatic procedure to analyze real-time magnetic resonance images (rt-MRI) of the human vocal tract recorded for read and spontaneously spoken speech. Significant differences were observed in the realizations of constriction-forming events for read and spontaneous speaking styles. Such an analysis has implications for understanding speech planning and for informing design of automatic speech analysis algorithm.

Reddy, S. and Yegnanarayana, B. Incorporation of Excitation Source and Duration Variations in Speech Synthesized at Different Speaking Rates.Paper presented at Speech Prosody Conference, Chicago, May 2010.

The effect of speaking rate on the excitation source is examined using instantaneous fundamental frequency (F0) and perceived loudness (η). The instantaneous F0 and η seem to increase in the case of normal to fast speech, where as they are speaker-specific for the case of normal to slow speech.The results are tested in perceptual experiments.

Significant parameters for the perception of speaking rate are discussed.

Rapp, K. (1971). A study of syllable timing. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)*, 1/1971, 14-19.

Siivola, V. (2002)A survey of methods for the synthesis of the singing voice. Unpublished Manuscript.

In this paper, different methods for synthesis of the singing voice are studied. Of the models presented here, physical models have probably the strongest theoretical background. The spectral models try to generate a natural sounding voice by mimicking the spectrum of a real voice. In this paper, spectral modeling with vocoders, formant synthesizers, formant wave function synthesizers and frequency modulation are presented. The most natural sounding source for human singing voice is a human singer. One approach, common in speech synthesizers, is to make a database of real voices and use these to synthesize the voice. Since building a big enough database is not feasible, the system must be able to perform transformations to fill the missing slots in the database. In this paper, concatenative synthesis with pitch synchronous overlap add, analysis by synthesis/overlap add and excitation plus resonances are studied.

Stone, R. E., Cleveland, T. F., Sundberg, J., & Prokop, J. (2002). Aerodynamic and acoustical measures of speech, operatic, and Broadway vocal styles in a professional female singer. TMH-QPSR, 43(1), 017-029.

In an attempt to quantatively analyze the different singing styles, this paper examines a female subject with professional experience in both operatic and Broadway styles of singing, who sang examples in these two styles.
How representative the examples are of the respective styles was investigated by means of a listening test. Further, as a reference point, the styles of her speech was also compared. Variation in styles associated with pitch and vocal loudness was investigated for various parameters: subglottal pressure, closed quotient, glottal leakage, H1 -H2 difference (the level difference between the two lowest partials of the source spectrum), and glottal compliance (the ratio between the air volume displaced in a glottal pulse and the subglottal pressure). Formant frequencies, long-term-average spectrum and vibrato

characteristics were also studied. Characteristics of operatic style emerge as distinctly different from Broadway style, the latter being more similar to speaking.

Strangert, E. (2002). Boundaries and groupings - the structuring of speech in different communicative situations: a description of the GROG project.. Proceedings of Fonetik, TMH-QPSR, 44(1), 065-068.

The goal of the project is to model the prosodic structuring of speech in terms of boundaries and groupings. The modeling will include different communicative situations and be based on existing as well as new speech corpora. Production and perception studies will be used in parallel with automatic methods developed for analysis, modeling and prediction of prosody. The model will be perceptually evaluated using synthetic speech.

Strangert, E. (2007). What makes a good speaker? Subjective ratings and acoustic measurements. Proceedings of Fonetik, TMH-QPSR, 50(1), 29-32. [pdf]

The paper deals with qualities contributing to the impression of a "good speaker"– a speaker capable of catching the attention of an audience through her/his way of speaking. Subjective ratings of speaker qualities were correlated with acoustic analyses of samples of speech produced in Swedish parliament debates. Raters reliably differentiated between more and less skilled speakers and reached good agreement on qualities contributing to their impression. Acoustic measurements reveal substantial differences with regard to F0 and duration features for speakers rated high and low, respectively, on speaker skill.

Sundberg, J., 1987. *The Science of the Singing Voice.* Illinois: Northern Illinois Press.

Sundberg, J., 1989. Synthesis of Singing by rule. In Mathews, M. V. and Pierce, J. R.

[Eds], ***Current Directions in Computer Music Research.***

***Sundberg, J. (2003). Research on the singing voice in retrospect. TMH-QPSR, 45(1), 011-022.**

A overview of the research on singing from the acoustic point of view. Sundberg's research on singer's formant, formant tuning, breathing, and voice source are reviewed in the light of later contributions. Detweiler's and Wang's studies of the singer's formant are commented. The idea that singers tend to tune F1 and/or F2 to harmonic partials is analysed and some open questions are pointed out. Various investigations of the voice source and breathing are discussed and some attractive topics for future research are described.

***Sundberg, J. (2006). "The KTH Synthesis of Singing." in Advances in Cognitive Psychology, 2006.v2,no.2-3,pp.131-143.**

This is an overview of the work with synthesizing singing that has been carried out at the Speech Music Hearing Department, KTH since 1977.

***Sundberg, J., Trovén, M., & Richter, B. (2007). Sopranos with a singer's formant? Historical, Physiological, and Acoustical Aspects of Castrato Singing. QPSR, 49, 1-6.**

Focusing on castrato singing,this paper combines the information embedded in historical sources, and the knowledge of the development of the human vocal organs, the aim being to explore hypotheses regarding the acoustical properties. The basic assumption is that the castrato voice combined the male adult vocal tract with the prepubertal voice source. A well trained boy soprano's rendering of Franz Schubert's Ave Maria on the vowel /a/ was inverse filtered and the voice source thus obtained was processed by a vocal tract filter with formant frequencies adjusted in such a way that a singer's formant with a centre frequency corresponding to an operatic tenor, baritone and bass voice was

obtained.

***Thalén, M., & Sundberg, J. (2000). A method for describing different styles of singing. A comparison of a female singer's voice source in "classical", "pop", "jazz" and "blues". TMH-QPSR, 41(1), 045-054.**

This investigation attempts to describe voice source differences between classical, pop, jazz and blues styles of singing as produced by a professional female singer and voice pedagogue at the pitches A3, C#4, E4 and G4 in soft middle and loud phonation. The voice source was analyzed by inverse filtering the flow signal. Four parameters were considered: (1) subglottal pressure, captured as the oral pressure during p-occlusion; (2) closed quotient; (3) the level difference between the two lowest source spectrum partials; and (4) the glottal compliance, defined as the ratio between the air volume contained in a voice pulse divided by the underlying subglottal pressure.

Van de Weijer, J. (2002). Terminal rises in infant-directed and adult-directed questions. Proceedings of Fonetik, TMH-QPSR, 44(1), 005-008.

Questions addressed to a prelinguistic infant do not have the same function as those addressed to an adult.This paper investigated whether, as a consequence of this differing function, infant-directed questions have a terminal rise equally often as adult-directed questions. The author analyzed a sample of infant-directed and adult- directed Yes-No questions, Wh questions, and, as a control, statements. The results showed no significant difference between the infant-directed and the adult-directed statements. On the contrary, the infant-directed questions ended with a terminal rise significantly less often than the adult-directed questions.

Wu, C. and Shih, C. Articulatory Effort in Different Speaking Rates.Paper presented at Speech Prosody Conference, Chicago, May 2010.

This study demonstrates articulatory effort at different speaking rates by examining articulatory trajectory using the Electromagnetic Articulograph AG500. The results suggest that the articulator undershoots a target and the valley of the target might go deeper while the velocity increases.

***Wu, W.H. (2009).An Acoustic Study of The Singer's formant: The comparison Between Western Classical and Traditional Chinese Opera Singing Techniques.Ph.D dissertation, Indiana University.**

The singer's formant (Fs) is a prominent spectrum envelope peak near 3000 Hz that appears in voices sung by trained Western classical singers. It is a raising cluster of formant 3, 4, and 5 and is especially important since this energy allows singers voices to be heard over the loud orchestra in the big concert or opera (Fant, 1970, Sundberg, 1970). Over the years, numerous researches have investigated the Fs by using many different methodologies. This study was to investigate the Fs by comparing two completely different training techniques: Trained Chinese opera singing techniques vs. Western classically trained singing techniques.

***Xu, Y. 2004. Understanding tone from the perspective of production and perception. Language and Linguistics 5: 757-797.**

This study addresses the question of discrepancy between speech production and perception. The maximum speed of pitch change and the coordination of laryngeal and supralaryngeal movements impose certain impassable limits on the way lexical tones are produced. At the same time, although the human perceptual system is highly proficient in processing fast-changing acoustic events as well as resolving distortions due to articulatory constraints, there are limits as to how much undershoot can be perceptually reversed. The understanding of these constraints has led to the Target Approximation

model of tone production.

***Xu, Y. and Sun X. 2002. Maximum speed of pitch change and how it may relate to speech. Journal of the Acoustical Society of America 111: 1399-1413.**

In the present study of constraints of speech production, a new experimental paradigm was adopted in which subjects produced rapid successions of pitch shifts by imitating synthesized model pitch undulation patterns(English and Mandarin Chinese). Results show that excursion time is nearly twice as long as response time. This suggests that physiological limitation on the speed of pitch movement is greater than has been recognized. Various implications of the experiments are discussed.

***Xu, Y. Patel, A. Wang, B. 2010. The role of F0 variation in the intelligibility of Mandarin sentences. Speech Prosody 5th International Conference, Chicago, IL.**

This study tested the importance of F0 variation for tone language comprehension. The intelligibility of\ Mandarin sentences with natural F0 contours was compared to the intelligibility of monotone (flat-F0) sentences created via speech resynthesis. In a quiet background, flat-F0 speech was just as intelligible as natural speech (about 94% intelligible), highlighting the robustness of the language comprehension system. However, when babble noise was added (0 db SNR) flat-F0 speech was substantially less intelligible than natural speech (60% vs. 80% intelligible), indicating that F0 variation is very important for Mandarin sentence intelligibility in noise.

Yoon, T.J. Speaker consistency in the realization of prosodic prominence in the Boston University Radio Speech Corpus.Paper presented at Speech Prosody Conference, Chicago, May 2010.

An analysis is presented on the rate of inter-speaker consistency in the way multiple speakers realize prosodic events when they read the same scripts. The results indicate that the average rate of consistency on the presence or absence of pitch accent is 89.81%. An average consistency of 72.17% is achieved for the rate of consistency for the types of the pitch accent. The finding implies that there is a constraint that is imposed on an utterance by speakers regarding prosodic prominence placement, as well as certain degree of variation between speakers in rendering prosodic prominence.

***Yoshida, K; Iversen, J; Patel, A; Mazuka, R; Nito, H; Gervain, J; and Werker, J. The development of perceptual grouping biases in infancy:A Japanese-English cross-linguistic study. Cognition 115 (2010) 356–361.**

Perceptual grouping has traditionally been thought to be governed by innate, universal principles. However, recent work has found differences in Japanese and English speakers' non-linguistic perceptual grouping, implicating language in non-linguistic perceptual processes (Iversen, Patel, & Ohgushi, 2008). Two experiments test Japanese- and English-learning infants of 5–6 and 7–8 months of age to explore the development of grouping preferences. The results reveal an early difference in non-linguistic perception between infants growing up in different language environments. The possibility that infants' linguistic phrasal grouping is footstrapped by abstract perceptual principles is discussed.

Zhang, S. Speech-to-song Illusion: Evidence from MC. Paper presented at SELLC 2010, Sun Yat-sen University/University of Helsinki, Guangzhou, China, 2010.

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.