On Words and Sounds

On Words and Sounds:
A Selection of Papers from the 40th PLM, 2009

Edited by

Kamila Dębowska-Kozłowska
and Katarzyna Dziubalska-Kołaczyk

CAMBRIDGE
SCHOLARS
PUBLISHING

On Words and Sounds:
A Selection of Papers from the 40th PLM, 2009,
Edited by Kamila Dębowska-Kozłowska and Katarzyna Dziubalska-Kołaczyk

This book first published 2011

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-3161-1, ISBN (13): 978-1-4438-3161-1

TABLE OF CONTENTS

PREFACE

The present volume *On Words and Sounds* is a collection of selected papers from PLM2009. The Poznań Linguistic Meeting (PLM) is an annual general linguistics conference that continues the tradition of the Polish-English contrastive conferences started by Jacek Fisiak in 1970. The new name "Poznań Linguistic Meeting" and profile were introduced in 1997 by Katarzyna Dziubalska-Kołaczyk when she took over as the Head of the Organising Committee. The Meetings are organised by the School of English, Adam Mickiewicz University, Poznań.

The book consists of fifteen articles, each of which can be read separately or in relation to others. The book will definitely appeal to the academic readership interested in the linguistic disciplines such as: phonetics and phonology, morphology, syntax, sociolinguistics, pragmatics and clinical linguistics. Collectively, the contributions investigate the interrelationships among those disciplines as well as between language and music. The central aim for the scholars was to explore PLM2009 leitmotif 'Variants, Variability, Variation' and show that the complete study of language involves diversified frameworks often rooted in the interdisciplinary approaches. This book aims at bringing together scholars aware of the need to merge individual linguistic disciplines and to provide comprehensive models for the study of the complexity of language.

—Kamila Dębowska-Kozłowska and Katarzyna Dziubalska-Kołaczyk,
Editors

The Speech-to-Song-Illusion Revisited

SIMONE FALK
LUDWIG-MAXIMILIANS-UNIVERSITY, MUNICH, GERMANY
AND TAMARA RATHCKE
UNIVERSITY OF GLASGOW, UNITED KINGDOM

Abstract

The present study investigates the boundaries of speech and song from an acoustic-perceptual perspective. Using the speech-to-song illusion as a method, we tested rhythmic and tonal hypotheses in order to find out whether acoustic characteristics can cue the perceptual classification of a sentence by German listeners as sung or spoken. First, our results show that, despite individual differences, the speech-to-song illusion is a robust perceptual phenomenon comparable to those known in visual perception. Second, the experiment revealed that 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.

1. Introduction

The question whether music and language should be considered as modular entities or share common resources and processes has been hotly debated in recent discussions (Peretz and Coltheart 2003, Patel 2008, Peretz in press). Concerning brain modularity, the phenomenon of song is especially intriguing as it obviously combines musical and linguistic structures. The existence of musical speech or music with words seems *per se* to corroborate the hypothesis of at least some shared domains (Patel in press). Nevertheless, the idea that singing might be exclusive to a language- or music-specific module and that during singing "music may act as a parasite of speaking" or vice versa has also been discussed (Peretz 2003, in press).

With respect to this controversy, it is worth taking a closer look at the mechanisms and acoustic premises guiding the perceptual conceptualisation of speech as opposed to song. In particular, we are interested in a perceptual phenomenon – the ‘speech-to-song-illusion’ (see Deutsch 1995, Deutsch et al. 2008) – that provides some evidence that an acoustic signal can be perceived both as speech and song.

1.1 An auditory illusion: speech-to-song

The illusion arises when a spoken utterance is presented in a loop so that the same prosodic structure with its speech-related characteristics is repeated over and over again. Surprisingly, listeners tend to perceive a shift from speech to song during the course of repetitions when they are told to judge the phrase as song- or speech-like. Unlike most visual illusions, this auditory effect seems to be unidirectional, i.e. once perceived as sung, a phrase cannot be perceived as spoken again (Deutsch 1995). So far, the illusion has only been studied with English listeners using a single English phrase (*(but they) sometimes behave so strangely*) which was originally spoken in context by the author herself (Deutsch 1995).

In an experiment by Deutsch et al. (2008), subjects listened 10 times to this phrase. After each repetition, they were asked to judge the phrase on a five-point-scale as speech or song. Unfortunately, the depiction in Deutsch et al.'s report only compares the judgments after the first and last presentation of the stimulus. This leaves the question open if the perceptual shift occurs abruptly at some point in the loop or if it is rather a continuous process. Additionally, it has been reported that the acoustics of the looped stimulus had to be unchanged since slight transposition of pitch or random permutation of syllables blocked the illusion (Deutsch et al. 2008). The speech-to-song illusion seriously challenges a strictly modular view of speech and music processing. In a modular architecture (Peretz and Coltheart 2003), the information available in the acoustic input is sent to task-specific modules that specialise in different sorts of input features for further processing. This implies that acoustic characteristics would predetermine if the incoming signal was processed in either music- or speech-related modules. The speech-to-song-illusion poses a problem since the acoustics of the same input can be perceived and processed both as speech and song without changing the incoming signal. Therefore, Deutsch et al. (2008) – while adhering to the modular concept – proposed that the same input activated separate pathways of speech versus song processing during the task. They also concluded that the perception of a phrase as speech or song could not be determined by specific music- or speech-like acoustic properties.

1.2 On the acoustics of speech vs. song

There are mainly two core features – pitch and rhythm – which are shared by both speech and music. However, these features are structured in a music/language-specific way, i.e. rhythm and pitch are associated with different acoustic instances depending on which medium/style they represent.

As far as pitch is concerned, scalar structure is said to be one of the most prominent musical features that speech generally lacks (Krumhansl 2005). Scalar structure in Western tonal music is represented by two scales, the major and the minor. For each of them, several tonal intervals are precisely defined in their tonal relationships and each deviation from these intervals is perceived as dissonant (Schellenberg and Trehub 1996). There is no such strict system of allowed tonal intervals in speech. The scaling of pitch targets and their intervals can vary considerably, the variation being conditioned phonetically (e.g. Lieberman and Pierrehumbert 1984), phonologically (e.g. Niebuhr 2007) or by discourse functions like grade of prominence (e.g. Ladd and Morton 1997). Furthermore, f0-trajectories of speech are characterised by f0-movements with variable slopes (t'Hart, Collier and Cohen 1990) whereas temporal stability of f0-targets is typical of pitch events in music (Zatorre et al. 2002). These differences are illustrated in Fig. 1 showing a German sentence (*Im Herbst bricht sich das Licht im bunten Laub,* English: 'In autumn, light is refracted by colourful leaves.'), which was both spoken and musically interpreted by the first author. As shown in the left-hand panel, the f0-trajectory of the sung phrase looks like a staircase leading up and down as it is formed by several f0-levels of equal intervals (seconds) whose number roughly corresponds to the number of syllables in the phrase. In contrast, f0-trajectories of the spoken phrase shown in the right-hand panel contain a series of falling and rising movements which vary in their slopes and intervals.

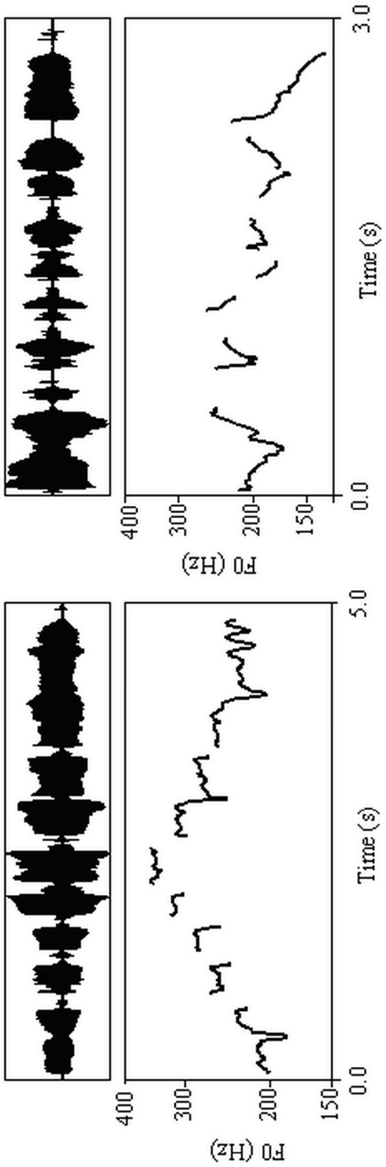

Figure 1: Waveforms and f0-trajectories of the sung (left-hand panel) and spoken (right-hand panel) German utterance “Im Herbst bricht sich das Licht im bunten Laub” (see text). F0 is scaled logarithmically.

Rhythm can be defined as the patterning of accentuation and grouping in sequences of events which systematically recur in time (e.g. Arvaniti 2009, Clarke 1999). As far as grouping is concerned, a composition of segments organised in syllables is widely agreed to be the smallest rhythmical unit of speech (Cutler 1991). In contrast, intervocalic intervals were assumed to constitute the smallest rhythmic units in song (Sundberg 1989). The predominance of intervocalic to syllabic chunking in song is demonstrated in Fig. 2, comparing both possibilities on the same sung German phrase *Schneeflöckchen, Weissröckchen, wann kommst du geschneit* (taken from a Christmas song; in English: 'Little snowflake, little white skirt, when will you appear to come down to earth'). As pointed out in Fig. 2, there is a tendency towards equal temporal spacing when segments of a sung phrase with equal note durations are arranged as intervocalic intervals but not as syllables: the mean of syllabic duration *dur (*S) is 0.47 sec with a standard deviation of 0.16 sec, whereas the mean of intervocalic intervals *dur(ii)* is 0.45 sec with a standard deviation of only 0.10 sec. Related to the temporal structure, isochrony has been widely discussed in both speech and music. However, speech patterns widely failed to show any isochrony in production (see Arvaniti 2009, Nolan and Asu 2009 for a critical overview), whereas temporal patterns in music are generally structured in a more systematic, isochronous way (metrically bounded music: e.g. Cooper and Meyer 1960, Drake and Palmer 1993).

Against the background of considerations given above, the English phrase used by Deutsch et al. (2008) was remarkable with respect to its tonal make-up. First of all, the phrase-final interval of 8 semitones (st) constitutes a minor sixth which is a well-defined consonant musical interval. Furthermore, besides some micro-prosodic f0-perturbations induced by consonants (Lövqvist et al. 1989), there are several relatively plain f0-stretches during the course of the sentence (especially the level-like production of the phrase-final word *strangely),* which rather relates to the sung than to the spoken example in Fig. 1. Hypothetically, the coincidence of relatively stable f0-trajectories and the harmonic interval in phrase-final position may have facilitated the occurrence of the speech-to-song illusion as reported in Deutsch et al. (2008).

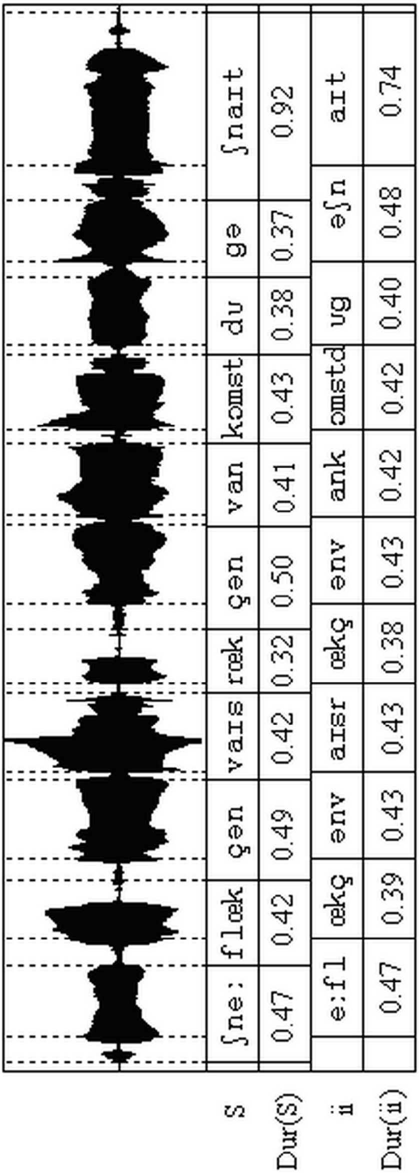

Figure 2: Labels and durations of syllables (S) as opposed to intervocalic intervals (ii) of the German phrase *Schneeflöckchen, Weissröckchen, wann kommst du geschneit* sung by an amateur female singer.

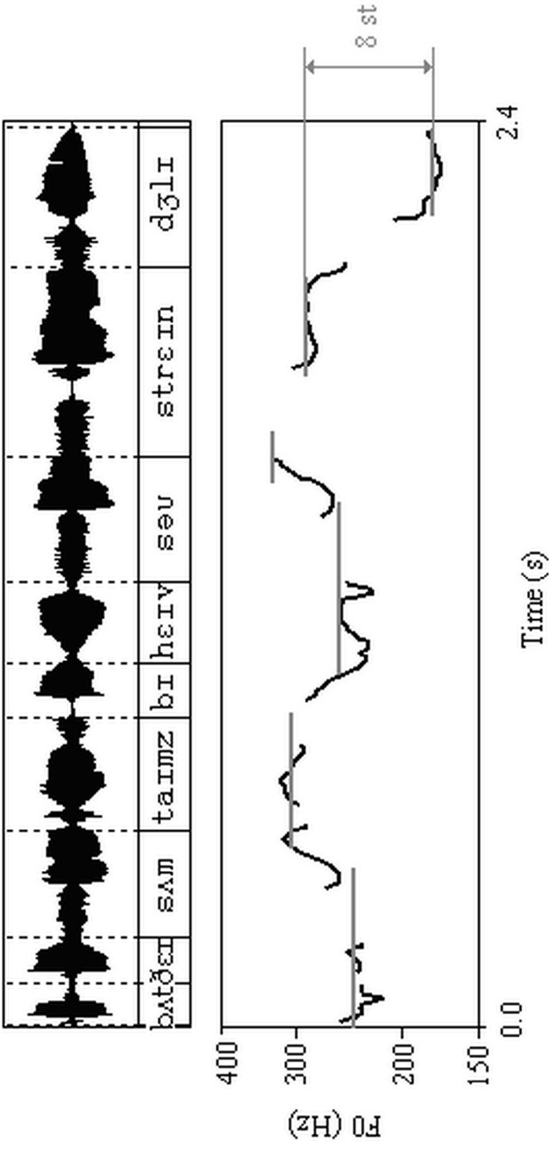

Figure 3: Waveform, syllabic labelling and f0-trajectory of the English phrase *but they sometimes behave so strangely* used in the illusion experiment by Deutsch et al. (2008). See text for more details.

1.3 Musicality: external factors of illusion perception

As we know from speech and music pathology, there is a certain autonomy between song on the one hand and speech perception and production on the other. For one thing, there are aphasic subjects with severe expressive speech pathology who are still able to sing (Warren et al. 2003, Mogharbel et al. 2005/2006, Peretz et al. 2004a). On the other hand, there are acquired or congenital amusics who can perform quite normally in speech-prosody related tasks (Peretz and Hyde 2003, Hyde and Peretz 2004). Congenital amusia is a multiple-faceted music processing deficit which affects the ability to, for example, judge if two melodies are the same or different, detect when music is out of key, produce pitch intervals and/or recognize what should be familiar tunes from their culture (Ayotte et al. 2002, Dalla Bella et al. 2009). The core deficit in this disorder concerns pitch processing, therefore individuals with this deficit are also called tone-deaf in the literature (Foxton et al. 2004, Hyde and Peretz 2004).

However, a deeper link between the processing of melodic and intonational pitch is also discussed (Patel et al. 1998, Nicholson et al. 2003). Patel et al. (1998) studied two individuals with acquired amusia subsequent to cortical brain damage. The salient finding from this study was that for both amusics, performance in a linguistic intonation task was very similar to performance in a tone sequence task, suggesting shared processing of melodic contours across the two domains.

Besides pathological cases, there is some evidence that music education can positively influence the subjects’ ability to differentiate intervals of varying range (see t’Hart 1981). For these reasons, we decided (1) to take into consideration musical training and (2) to perform a subtest of the Montreal Battery of Evaluation of Amusia (MBEA, Peretz et al. 2003) in order to assess whether some subjects could be at risk to be congenitally tone-deaf.

2. Perception experiment

2.1 Hypotheses

In contrast to Deutsch et al. (2008), we generally assume that acoustic properties of a given signal do play a role in inducing a perceptual shift from speech to song (the *Main Hypothesis*). As discussed in 1.2, the stimulus used for the original illusion task has inherently been optimal in its tonal layout to generate the effect. Therefore, we are going to explore whether and which acoustic features of a stimulus are better than others in inducing the shift from speech to song. More precisely, we hypothesise that two sets of signal-related features can account for (1) the fact that the auditory illusion emerges and (2) how rapidly it occurs within a loop.

The *Tonal Hypothesis* assumes that the perceptual shift is predominantly induced by tonal properties of a spoken utterance, namely:

(1.a) *Target stability:* If the tonal make-up of a sentence has more stable tonal targets, the shift will occur earlier compared to a sentence with unstable targets.

(1.b) *Interval structure:* If the tonal properties of a sentence involve scalar relationships of music, the shift will occur more easily.

In contrast, the *Rhythmic Hypothesis* assumes that the shift is primarily facilitated by rhythmic factors such as:

(2.a) *Accentual regularity:* If accented syllables are more regularly spaced, the shift will occur earlier compared to a non-regular distribution of accents since temporal regularity of beats increases the perception of a strong rhythm.

(2.b) *Segmental grouping:* Temporal grouping of segments into intervocalic intervals of same length will lead to an earlier shift from speech to song in contrast to a syllabic grouping of segments.

2.2 Method

2.2.1 Materials and test stimuli

With respect to the *Rhythmic Hypothesis* (2.a), two sentences were chosen as experimental materials (***bold*** = accented):

*Im Re**gal** liegen **Na**del und **Fa**den.* (English: 'There are needle and twine on the shelf.')
*Im **Gar**ten blühen heute **Klee** und **Mohn**.* (English: 'Today, clover and poppy are blooming in the garden.')

Spoken with broad focus, accents appear regularly in the first sentence, constituting a kind of anapest (two weak units are followed by a strong beat) as opposed to the second sentence where there is no such regularity. However, both sentences have 10 syllables and 3 accents in a broad focus.

Four sentences were added to serve as fillers (see 2.2.2). All sentences were read by a female speaker of Standard German and recorded in a sound-isolated booth at the Institute of Phonetics and Speech Processing in

Munich. All durational and f0-manipulations were done using *Praat* (Boersma and Weenink 2001). Typical f0-values of the speaker as well as her intonation patterns produced in both test utterances were used as the basis for creating tonal stimuli. To implement the tonal target stability hypothesis (1.a), f0-contours between relevant f0-targets were either kept stable or changed gradually with respect to segmental landmarks. Regarding the tonal interval structure hypothesis (1.b), we implemented the perfect fifth (interval of 7 semitones, st) twice in the signal: once as an ascending interval at the first pitch-accent and again as a descending interval at the last pitch-accent of the sentence. The interval in the speech-like condition was set to 5.5 st which should not be interpreted in terms of music scale. A top and base-line declination of 0.5 st was applied to pitch-accented and unaccented syllables. Fig. 4 compares f0-trajectories of stimuli created for both test sentences: (1) music-like stimuli with scalar structure and stable f0-targets are indicated by black lines, whereas (2) speech-like stimuli with non-scalar relationship between f0-targets and gradually changing f0-contours are shown by grey lines.

To test the temporal grouping hypothesis (2.b), the sentences had to be chunked into syllables as well as into intervocalic intervals. Due to schwa reduction in both test sentences (in /liːgən/, /blyːən/ and /faːdən/), distinguishing 10 rhythmic units for both conditions proved difficult. We decided to consider phrase-medial reduced forms as single rhythmic units, i.e. [liːgn], [blyːn] and phrase-final [faːdn̩] as two different units (see Tables 1 and 2). This decision was made with respect to the finding that phrase-final rhythmic units tend to have additional duration when compared to phrase-medial units (e.g. Lehiste 1973). As shown in Table 1, intervocalic intervals all had equal duration. In contrast, in the speech-like condition (Table 2), unaccented syllables were shorter than accented ones, as is usually the case in speech (e.g. Beckman and Edwards 1990). In each sentence, duration of the final rhythmic unit accounted for final lengthening. The total length of every test sentence was 1.85 sec. Note that there were some small deviations from these target values as manipulations were done manually.

The duration and tonal make-up of the filler stimuli were also manipulated in order to establish a homogenous experimental signal. Four rhythmic and/or tonal interpretations were created for each of the four fillers. The procedure described above resulted in 16 test stimuli (2 sentences x 2 durations x 2 f0-contours x 2 f0-ranges) and 16 filler stimuli (4 sentences x 4 interpretations). Each of the 32 stimuli was looped with 10 repetitions and a 0.4 sec pause between them. Note that the pause within the loop used in Deutsch et al. (2008) was twice as long as the pause implemented in our stimuli. During pilot sessions, a shorter pause seemed to support the illusion effect.

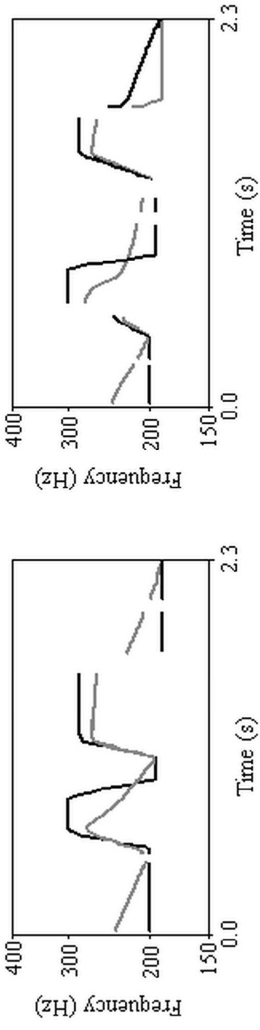

Figure 4: F0-trajectories of the music-like tonal structure (black lines) and the speech-like tonal structure (grey lines) in the test sentence *Im Regal liegen Nadel und Faden* (left-hand side) as opposed to the test sentence *Im Garten blühen heute Klee und Mohn* (right-hand side).

Table 1. Chunking and duration of intervocalic units (in sec, song-like condition 2.b). Capitals: pitch-accented vowels.

Units	imr	eg	All	ignn	Ad	el	undf	Ad	n	Total
	img	Art	nbl	ünh	eut	ekl	E	undm	On	
Duration	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.25	1.85

Table 2. Chunking and duration of syllabic units (in sec, speech-like condition 2.b). Capitals: pitch-accented vowels.

Units	im	re	gAl	lign	nA	del	und	fA	dn	Total
Duration	0.15	0.15	0.25	0.15	0.25	0.15	0.15	0.35	0.25	1.85
Units	im	gAr	tn	blün	heu	te	klE	und	mOn	Total
Duration	0.15	0.25	0.15	0.15	0.15	0.15	0.25	0.15	0.45	1.85

2.2.2 Filler stimuli

The fillers were pairs of sentences with syntactic and rhythmic structure similar to the target test sentences, but they differed from those semantically, morpho-phonologically and partly with respect to the number of pitch accents. The following sentences served as fillers (***bold*** = accented):

> *Die **Rin**der kann man nicht **fin**den.* (English: 'You cannot find the cows.')
> *Die **Kin**der soll man nicht **bin**den.* (English: 'You should not take possession of your children.')
> *Im **Him**mel **sin**gen die **En**gel.* (English: 'In heaven, angels are singing')
> *Im **Stall quen**gelt der **Ben**gel.* (English: 'In the barn, the baby boy is whining')

The fillers were manipulated both tonally and rhythmically resulting in 16 additional stimuli (4 sentences x 4 manipulations). Rhythmical manipulations were applied randomly to the four sentences and comprised varying lengthening grades of intervocalic intervals, syllables or whole phrases. For each sentence, two stimuli with level-like and two stimuli with contour-like tonal make-ups were included.

2.2.3 Procedures

The total of 32 stimuli was divided into two sets, each containing 8 test and 8 filler sentences in order to keep test duration as short as possible and to prevent subjects from hearing sentences with the same semantic content too often, i.e. no more than four times. Each test started with three practice sentences among which we presented the original English stimulus (Deutsch 1995) and two German practice sentences. For the test phase, 16 stimuli were pseudo-randomised using one randomisation list per subject. Randomisation was guided by the following rules: (1) the test always started with a filler sentence and (2) between two test sentences with the same semantic content, at least two other (test or filler) sentences always intervened.

The subjects' task was to listen to the looped stimuli and to press a button as soon as they had the impression that the signal was no longer a spoken phrase, but had song or music-like qualities. Interestingly, we observed an effect the instruction had on the perception of the illusion during pilot sessions: if subjects' attention was not called to the potential shift from an auditory impression of a spoken phrase to that of a sung phrase, they tended to report cases of verbal transformations (Warren 1961, see section 3 of this paper) or perceived some changes in prosodic structure of the looped stimuli. At this point, formation of the illusion reflects the main perception principle attested for both visual and auditory modalities being dependent on expectation and experience (see Hawkins 2009 for an overview).

After each trial, subjects had to confirm whether they really had perceived a change in the signal or not. Subsequently, they were asked to solve a simple mathematical equation. This was done to interrupt the test phase, to distract subjects' attention and to clear working memory load from sound imprints induced by massive repetition of the previously heard stimulus. After the main test phase, subjects additionally performed the scale subtest of the MBEA (Montreal Battery of Evaluation of Amusia, Peretz et al. 2003). In this test, subjects had to compare musical phrases (i.e. tone sequences that differ in one pitch relationship or not) and to judge them as "same" or "different". This test was included as we expected an influence of the subjects' musical expertise on their ability to perceive the illusion more easily. Additionally, we obtained self-reported musical capacity with scores ranging from 0 (non-musical at all) up to 10 (very musical).

The last task of the experimental session was to judge the quality of the non-repeated stimuli as spoken or sung (singing/speaking-style rating task). The subjects were asked to listen to the complementary set of test and filler sentences (only one presentation per stimulus) and to decide on a seven point scale whether stimuli were spontaneously perceived as clearly/ quite/ rather spoken (1-3) or rather/ quite/ clearly sung (5-7); unclear cases were to receive the label 4. All tests were run on computer using DMDX software (Forster and Forster 2003). Auditory signals were presented via Sennheiser head-phones of good quality. The session lasted about 25 min.

2.2.4 Subject sample

In total, 62 native speakers of German (13 m, 49 f) aged between 19 and 46 (mean age: 24 years) participated in the experiment. 52 subjects reported having some experience with music (singing or playing an instrument). On average, these subjects had started musical training at the age of 8 and continued for about 7 years. The rest of the sample (10 subjects) had no music education. Each subject was tested separately. The sample was divided into two groups (30 and 32 subjects). Each group was tested once using one of the stimuli sets (set 1 or set 2). The data for test stimuli from set 1 consisted of 30 judgments for each stimulus in the illusion task and 32 decisions in the singing/speaking-style rating task, whereas for set 2, there were 32 judgments for each stimulus in both tasks.

2.3 Results

2.3.1 Tonal versus rhythmic cues

Overall, in filler and test stimuli, 60 out of 62 subjects experienced a shift from speech to song. Regarding the test stimuli alone, 59 out of 62 perceived the shift. The perception of the shift seems to be quite robust: in the test stimuli, 41 subjects (66 %) reported a shift in more than 50 % of the items. Fig. 5 gives an overview of the results showing the percentage of perceived illusions dependent on the four experimental factors. As predicted by our *Main Hypothesis*, acoustic characteristics did influence the overall perception of the illusion. In favour of the *Tonal Hypothesis,* tonal cues – in particular target stability – had the strongest effect on the perception of the illusion: stimuli with stable targets and interval structure were more likely to evoke the illusion than stimuli with unstable targets and without scalar relationships. Surprisingly, rhythmic factors did not show any obvious effects. Stimuli with regular pitch-accent distribution induced slightly more occurrences of the illusion, whereas stimuli with equal duration of intervocalic intervals were more likely to suppress rather than to facilitate the illusion, which was in contrast to our prediction (see 2.b). Repeated measures ANOVA performed with four independent factors (*target stability*, *interval structure, accentual regularity, segmental grouping*) and dependent variable being the number of perception shifts revealed significant results only for *target stability* (F=8.5 and p<0.01) and for the interaction *target stability * accentual regularity* (F=12.6 and p<0.001). As shown in Fig. 6, the illusion was most often perceived in stimuli with stable targets and accentual regularity whereas it was least often reported for stimuli lacking these two characteristics.

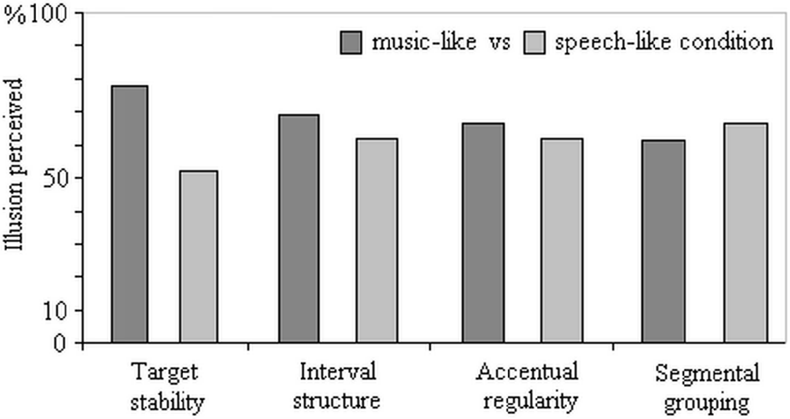

Figure 5: Percentage of perceived speech-to-song shifts dependent on the four experimental factors (averaged for two stimuli sets).

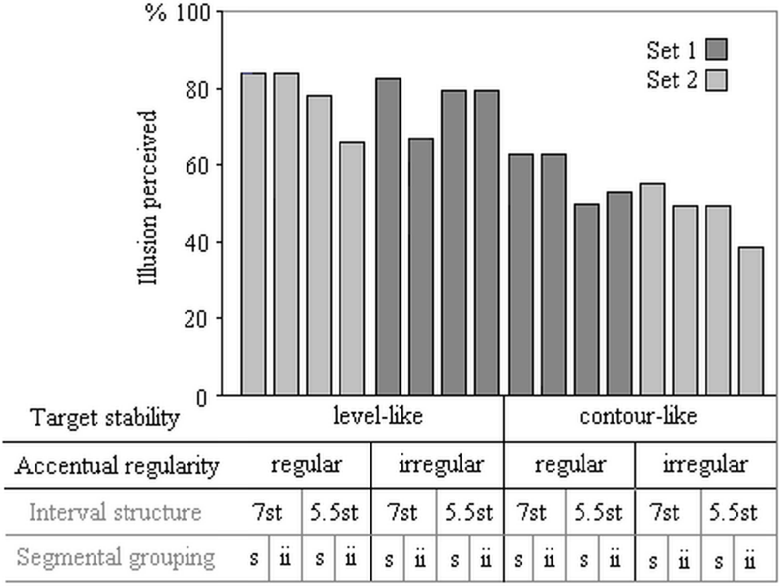

Figure 6: Percentage of perceived speech-to-song shifts for each test stimulus (n_1=30 for set 1, n_2=32 for set 2). Test factors are presented with respect to their potential to induce the illusion (in descending order from top to bottom).

Table 3 presents the frequency of perception shifts during the course of ten repetitions (1-10) compared to the frequency of failed illusion averaged for all stimuli. Note that a similar pattern of judgment distribution was observed for each experimental condition. According to repeated measures ANOVA, no significant differences were caused by the four tested factors, i.e. in contrast to our predictions, acoustic characteristics of the stimulus did not ‘slow down’ or ‘speed up’ the occurrence of the illusion, but rather contributed to the overall rate of perception shifts. Obviously, there was some variation in the speed of perception shift in our subject group (Median: 4th repetition). However, the illusion was mostly perceived during the third repetition of the stimulus in all conditions.

Table 3: Frequency of judgements (no illusion perceived vs. illusion reported during the 1-10 repetition) dependent on the four test factors (music-like *m-l* or speech-like *s-l*).

Factor		*No Ill.*	*Repetition*										*Sum*
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	
Target stability	m-l	55	11	30	54	33	28	10	11	8	5	3	248
	s-l	117	3	14	27	22	18	15	11	8	4	9	248
Interval structure	m-l	77	8	24	45	30	27	13	11	4	4	5	248
	s-l	95	6	20	36	25	19	12	11	12	5	7	248
Accentual regularity	m-l	79	9	26	46	24	26	8	11	9	5	5	248
	s-l	93	5	18	35	31	20	17	11	7	4	7	248
Temporal grouping	m-l	93	5	17	35	31	20	14	12	9	4	8	248
	s-l	79	9	27	46	24	26	11	10	7	5	4	248
Mean		*86*	*7*	*22*	*40.5*	*27.5*	*23*	*12.5*	*11*	*8*	*4.5*	*6*	*248*

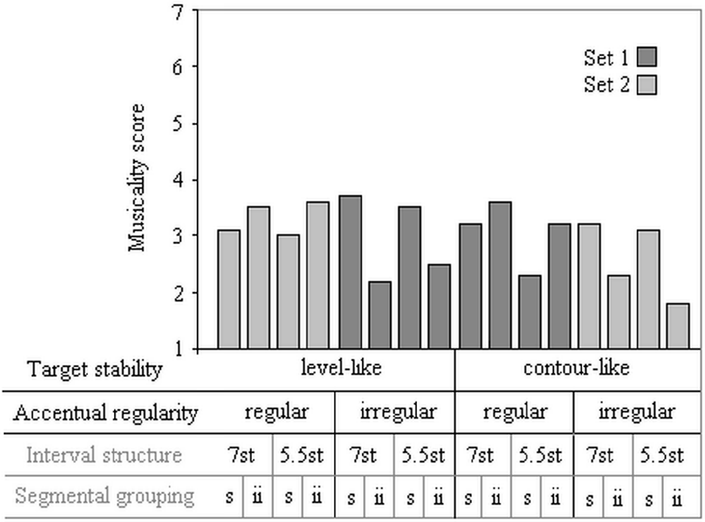

Figure 7: Mean scores for each test stimulus in the singing/speaking-style rating task (n_1=30 for set 1, n_2=32 for set 2). Test factors are presented with respect to their potential to induce the illusion (in descending order from top to bottom, as in Fig. 6).

Concerning the singing/speaking-style rating task (see Fig. 7), there is no correlation between the rating score of a stimulus and its potential to induce the illusion obtained in the illusion test. Obviously, perception of the illusion during the main test cannot be considered an artefact of stimuli acoustics, it being too song-like from the start. Overall, the test stimuli were judged as rather spoken (score range: 1.5-4, averaged score 3). Repeated measures ANOVA with four independent factors (*target stability, accentual regularity, interval structure, segmental grouping*) and dependent variable being the score from the singing/speaking style task revealed no significant effects for any of the test factors. Thus, we could assume that all our test stimuli are perceived more or less equal in their speech and song-like quality. One of the filler sentences received a relatively high score in the style rating task (*quite sung*: mean 5.2, median 6). This stimulus had stable targets and was lengthened with a factor of 1.6 to the original sentence (*Im Stall quengelt der Bengel*, see 2.2.2). Interestingly, this stimulus was not better in inducing the illusion (75% of cases) than the test stimuli with relatively low style rating scores (3-4 scale points, see Fig. 7; for overall illusion perception in 80-85% of the cases, see Fig. 6). In contrast, one test stimulus with intervocalic intervals, non-scalar structure, irregular accentual timing and non-stable targets received the lowest style rating score of about 1.5 and evoked the least number of perceptual shifts (under 40 %). In sum, we can assume that the illusion is more likely to occur as soon as the acoustics of a stimulus are somehow ambiguous between speech and song – as given in most of the test stimuli. Stimuli in which acoustics were too song-like did not facilitate perception of the illusion, but stimuli with too speech-like features did weaken the probability of the shift.

2.3.2 Individual perception patterns

The inter-individual ranking of stimuli presented above (Fig. 5) was also intra-individually consistent. Fig. 8 shows how often subjects (grouping of subjects was done by number of shifts reported in the eight test stimuli they heard) experienced the illusion on average in the two most often and two least often perceived stimuli per set. It can be seen that subjects with a low perception rate of the shift in the test stimuli (1 of 8, 2 of 8, 3 of 8) perceived it at least in one of the two stimuli that were perceived most often by the whole group of subjects. Additionally, in these subject groups, the shift was rarely perceived in the two least often perceived stimuli on the set. Furthermore, subjects who heard the illusion more often in the overall set of test stimuli, also heard it earlier in the looped test stimuli (n=62, negative correlation between mean number of illusion occurrences per subject and mean number of repetitions, R^2=-0.6, p<0.001).

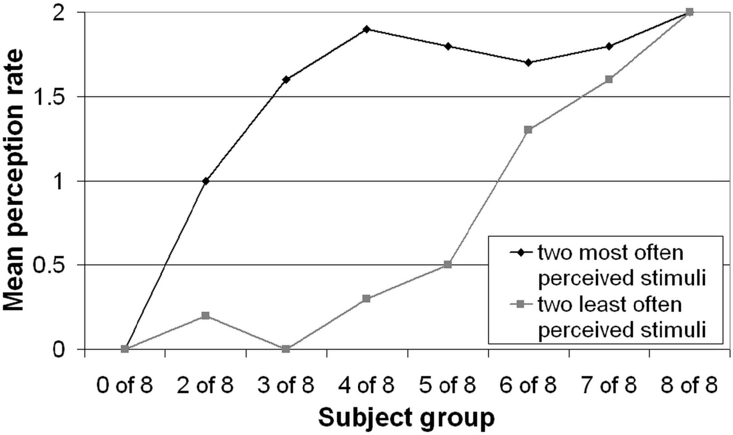

Figure 8: Mean perception rate of the two inter-individually most (black) and least (grey) often perceived stimuli.

Concerning musicality, averaged MBEA-values of subjects with versus without musical education differed significantly (p<0.01; t=3.25) showing that musicians were more sensitive to tonal changes (mean score 25 out of 30 points) than non-musicians (mean score 23 out of 30 points). However, there was no significant correlation between obtained MBEA-values and the number of illusion occurrences reported for the test stimuli nor did MBEA-scores correlate with the tendency to perceive the speech-to-song shift earlier or later in the loop. Overall, 7 subjects performed worse than the cut-off score (22) in the MBEA scale subtest, one performed on the level of the average of amusics as reported in Peretz et al. (2003). However, all of those subjects reported to hear the shift in 50 to 100 % of the test stimuli. Thus, we conclude that the scale subtest of the MBEA was not sensitive to predict the individual performance pattern of subjects in the illusion test.

3 Discussion

Our study confirms that the speech-to-song-illusion as described for an English sentence (Deutsch 1995, Deutsch et al. 2008) is robustly perceived by listeners of (at least) another intonation language, as we have demonstrated that the illusion emerges (1) in German and (2) with various sentences and manipulated speech.

In general, the results support our *Main Hypothesis*: acoustic properties of the signal did influence the perception of the illusion, and especially the overall frequency of its occurrence. So far, the study has mainly supported the *Tonal Hypothesis* since target stability significantly facilitated occurrence of the illusion. In contrast, the results were less clear in terms of the *Rhythmic Hypothesis*. We can assume that the latter's phonetic implementation was not powerful enough to yield prominent effects in this study. However, we cannot claim that our conception of rhythm was completely wrong as we observed a significant interaction between target stability and accentual regularity. Temporal regularity of accents seems to be a secondary cue which supports the perception of the illusion. Isochronous spacing between pitch accents – i.e. an exact "beat" structure – may induce an even stronger effect and should be tested in a further experiment.

Concerning the segmental grouping condition (2.b) we encountered some problems specific to German vowel reduction processes: test sentences with schwa-elision (see 1.4) led to difficulties in chunking of intervocalic intervals. As we know, vowels in speech differ from vowels in song depending on voice register, the latter exposing more peripheral qualities in lower registers that are also used in speech (Sundberg 1987, Falk 2007). Thus, vowel reduction or undershoot in speech could have been constraining perception of the illusion. Further testing with stimuli sentences lacking schwa-vowels will shed more light on this question.

The study has shown that there were individual differences in the perception of the illusion since some subjects heard the illusion more often (and sometimes earlier) in the loop than others. In general, intra-individual performance followed the inter-individual ranking of stimuli but the study did not reveal a basis for explanation of individual performance patterns. So far, we could not find a link between the performance pattern in the illusion test and performance in the MBEA scale subtest. A different battery for assessing musical abilities might be better suited to capture individual differences in the perception of the illusion.

Interestingly, when subjects did perceive a shift, they were most likely to experience it during the third repetition within the loop. This effect was also quite robust for all experimental conditions. As reported for some optical illusions (as e.g. Necker Cube or Rubin Vase, Gregory 1971)[1] or other auditory illusions like Auditory Stream Segregation (Bregman and Campbell 1971), a certain amount of time – or even a verbal instruction – is usually needed until re-interpretation of the originally perceived image/signal is possible and a perceptual shift occurs. We assume 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. This might be similar to production processes in a speech cycling task with the result "[...] that a higher level dynamic emerges within which the timing of subordinate processes are constrained" (Cummins and Port 1998).

In the verbal domain, several effects induced by massive repetition of verbal stimuli have been reported in the literature, i.e. "semantic satiation" (Smith 1984), "syntactic fatigue or satiation" (e.g. Francom 2009) and "verbal transformation" (e.g. Warren 1961, Ditzinger et al. 1997). These processes might share common grounds with the speech-to-song illusion as they create a situation in which the linguistic and grammatical meaning stepwise lose their importance. "Semantic satiation" is achieved by massive repetition of a word, during which subjects are experiencing a feeling of "detachment" of meaning during the course of repetition. In a spreading activation network model, semantic satiation is explained as an over-activation of the target word node which consequently fatigues and hinders the access of the meaning of the word. Thereby, phonological neighbour nodes in the lexicon become more and more available to the extent to which the target word node becomes satiated/ over-activated, leading to "verbal transformations". This term refers to the fact that subjects report hearing phonological similar words (e.g. *face, space, paste* instead of *pace*) during the repetition cycle (McKay et al. 1993). However, Pilotti et al. (1997) conducted an experiment showing that repetition of the same word in different speakers' voices blocked the satiation effect. The authors therefore conclude that the observed effect is less due to semantic satiation than to early acoustic adaptation to the signal (Eimas and Corbit 1973). This is an interesting parallel to the speech-to-song illusion where the variation of the tonal or rhythmic make-up of the given phrase within the loop blocked the illusion as well (cf. Deutsch 1995, Deutsch et al.

[1] Both Necker Cube and Rubin Vase are ambiguous drawings which allow two interpretations of foreground vs. background. A three dimensional wire-frame drawing of a cube does not show which is in front and which is behind when two lines cross. Similarly, a picture of a white hourglass-shaped vase against dark background is schematic and its shape is formed like two profiles facing each other.

2008). Another parallel to the findings of our experiment is that at least 3 to 4 repetitions are necessary before semantic satiation is observed (Pilotti et al. 1997). In our study the perceptual shift from speech to song also occurred most often during the third repetition. These parallels give rise to further research questions: Is semantic satiation a relevant process to the speech-to-song illusion or are both the effect of an earlier acoustic analysis process? How will differences in semantic and syntactic complexity affect the perceptual shift from speech to song? If there is underlying semantic satiation in the speech-to-song illusion, can we observe also the reverse process of song-to-speech by re-semantisation e.g. by instruction to concentrate on aspects of the meaning of the sentence?

Finally, our results do not allow direct conclusions about the modular or non-modular nature of speech and song processing. However, the fact that some acoustic properties of the signal in our experiment facilitated song perception better than others could point to a decoding strategy relying on specific acoustic cues to rate the "song-likeness" or "musicality" of a signal. Furthermore, it seems plausible that the decoding includes a re-evaluation of the acoustic structure at early stages of processing. 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. After the overall acoustic pattern has been "carved out" by the listener, it is free to be re-evaluated as a plausible or non-plausible musical event. Some aspects of the acoustics – as target stability – seem to be more powerful to cue musical interpretation which in the end leads to the perception of song instead of speech. This points in the direction of a modular approach, but more empirical evidence is needed to explore this issue.

4 Conclusions and open questions

The present study has shed some light on the processes underlying the speech-to-song illusion. First of all, we have demonstrated that the perceptual shift was robustly perceived by a group of German listeners in a variety of sentences and that the occurrence of the shift was influenced by the acoustic make-up of the stimuli. The results suggest that a certain ambiguity in the acoustics of a signal may be necessary to induce the shift. The most potent cue to the speech-to-song illusion in this experiment was tonal structure (stability of tonal targets), but we also found a certain potential of rhythmic properties to influence the perceptual shift. This is an issue that will be investigated in further experiments by exploring in more