Prosodic Alignments in Shadowed Singing of Familiar and Novel Music

Eran Raveh[1], Maya Twig[2], Bernd Möbius[1], Oded Zehavi[3]

[1]Language Science and Technology, Saarland University, Saarbrücken, Germany
[2]Musicology Department, Hebrew University of Jerusalem, Israel
[3]Department of Music, University of Haifa, Israel

raveh@coli.uni-saarland.de

Abstract

This paper presents a study comprising two singing shadowing tasks focusing on prosodic features of music. The first experiment investigated alignment effects in a song known to the participants. They sang the song before and after listening to a recorded version of it. The second experiment tested which prosodic elements are best preserved in replications of an unfamiliar song. Methods used in phonetic accommodation studies were adapted and used to measure the effects. Results show that convergence occurs in singing, but not in the same manner across all tested features. Additionally, participants preserved rhythmic patterns better than the tonal contour in the unfamiliar music piece.

Index Terms: music and speech, vocal accommodation, prosodic alignment, shadowing experiment

1. Introduction

The rhetorical aspects of music and spoken language can be described in musical terms. Melody, pitch, timbre, rhythm and tempo are all common in description of music, as they are in speech [1, 2]. Some spoken language related phenomena can also be described using musical means [as in 3].

In this paper, we investigate a speech-related phenomenon, namely *phonetic convergence*, in sung music. Phonetic convergence is a process in which interlocutors become more similar to each other in terms of phonetic features while interacting [4, 5]. Methods commonly utilized in this research topic were adapted to suit musical material to examine whether convergence can be found also in singing and whether the musical nature of the used materials can shed light on additional aspects of convergence.

The two types of vocal capabilities, singing and speech, share some properties in both production and perception. Such common properties include articulation rate, intensity, timbre, and others. Another important aspect is that both have a temporal dimension and evolve over time. However, music consists of defined *absolute* pitch and rhythmic targets. This is even more salient when dealing with an already known musical material, as both the singer and the listener already have expectations regarding the tones before they are produced [6]. In speech, on the other hand, the phonetic features of a specific utterance are not expected to match specific absolute values.

Phonetic convergence has been studied with respect to various prosodic features, such as speech rate [7, 8], fundamental frequency [9, 10], intonation [11, 12], rhythm [13], and more. Due to similarities and dissimilarities between speech and music, we believe it can be rewarding to see whether convergence can be found in singing as well, and whether it operates differently. As both speaking and singing are used in social contexts, external factors may potentially affect them. In the case of

Snippet 1: *The Yakinton lullaby transposed to B major. The square labels "A", "B", and "C" mark the* theme*, bridge (or* development*), and* recapitulation *sections of the song. The breath marks are placed where the participants are expected to make a brief break and/or lengthen the ending of a phrase. The first sixteenth note in bar six is in brackets since it is not present in the original melody and was therefore also excluded in the recorded version played to the participants. However, it is common to add it, and indeed all participants included it in both performances.*

music, convergence can be expressed in different aspects than in speech, e.g., singing the notes more accurately, shifting the musical key, adapting to a different tempo, etc. Furthermore, seeing that the tonal targets in singing are pre-defined, there are precise targets for production – either from a heard example or from one's mental memory – in an experimental setting. Therefore, the participants' productions can be directly compared with some "ground truth".

The main research question of this work is whether convergence occurs in singing as well, and, if so, whether specific parts of the musical piece are prone to undergo changes. Our hypothesis is that convergence will be found in the participants' performances with respect to the pre-recorded stimuli. This can be realized on the absolute level, meaning that the participants shift their overall pitch range (i.e., the *key*) and tempo to be closer to the recording, or relative to their own singing by making the pitch and temporal intervals between the target notes more accurate. An additional question is how the familiarity with the musical material affects reproduction. The expectation here is that the participants' performances of the familiar song will be accurate even before listening to a version of it in terms of deviation from the target intervals, but even more so after listening. When reproducing an unfamiliar melody, it is not ex-

Snippet 2: *Neutral lullaby. The square labels "A" and "B" mark the structural parts. The grace notes in bars 2, 6, and 12 were included in the recording but due to their secondary melodic role did not penalize performances that lacked them.*

pected that the participant will remember it in its entirety, but rather that they would stick to repeating segments or parts with smaller intervals and simpler rhythms.

The present study consists of two shadowing experiments [cf. 14]. Sung lullabies were chosen in this work, as they are presumably more memorable than other musical genres and than instrumental pieces [15, 16]. Shadowing – here, delayed, non-immediate shadowing – is a common methodology in convergence studies [e.g., 17]. The first experiment tests convergence effects between two performances of the participants, one before and one after listening to a pre-recorded version of a song *familiar* to them. In the second experiment, participants listened to another, *unfamiliar* song and replicated it to their best ability. This experiment tests which prosodic features would be replicated best when it is not likely that the participants would remember all aspects of the musical material. The experimental procedures and materials are described in Section 2. Section 3 presents the methods applied to analyze the prosodic properties *pitch*, *tempo*, and *rhythm*. Results are presented in Section 4, along with some representative examples from the participants' performances. Finally, interpretations and additional observations are discussed in Section 5, which also concludes the paper.

2. Method

2.1. Material

Two lullabies were used in the experiment: The first is "Tune for the Yakinton"[1] (hereafter "Yakinton", see Snippet 1), which is a famous Israeli children song. The second is a neutral lullaby composed for experimental purposes [18, pp. 22-47, see Snippet 2], which contains *cross-cultural similarities* such as repetitiveness, simple melody, and a limited inventory of tonal and rhythmic elements [19, 20]. Therefore, while the first one is assumed to be known to the participants, they could not be familiar with the second one. Both lullabies are short (13 bars at $\quarternote = 61$ ($\approx 26.5\,$s) and 16 bars at $\dottedquarter = 33$ ($\approx 58\,$s), respectively) and in major keys. Both songs were recorded a cappella by a trained female singer in the same age group as the participants in a professional recording studio at 44.1 kHz sampling rate and 16 bit resolution. To avoid changes in voice production, decrease vibrato, and limit the singing effort, the songs were transposed

---

[1]Pizmon LaYakinton, written by Leah Goldberg in 1940. *Yakinton* is the Hebrew name of the Hyacinth plant.

and recorded in B major, which is relatively low for female voices. This also avoids influences originating from the use of different keys. The syllable [na] was used throughout the songs in both the recordings and participants' productions to eliminate any bias due to their lyrics (see Section 2.3). This way, the possibility that the participants would know the melodies but not the lyrics was avoided as well.

2.2. Participants

Six participants, all of which are mothers with no hearing impairments to recently born babies, took part in the experiments. Their age ranged from 29 to 37 years (mean 35.5 ±3.25) and the age of their babies ranged from one to seven months (mean 4.5 ±3.5). For three of the mothers this was the first child. To further homogenize the participants' characteristics, their musical education and experience was controlled as well. None had any professional-level musical background, while four disclosed they have been playing an instrument or singing.

In the pre-verbal phase, parents often sing to their babies. When communicating with infants, adults tend to use exaggerated prosody with elevated melodic pitch and distinct rhythmic patterns [21]. The increased use of singing as well as its function as a means of communication with their babies [see 22, 23] made mothers of small babies suitable for this experiment. Since the participants' singing capabilities are essential for their performance in the experiment, it was also confirmed that they regularly sing lullabies for their babies. Furthermore, as we are dealing with a specific lullaby (see Section 2.1), their familiarity with it was also considered. All the participants reported that they know the familiar lullaby well enough to spontaneously sing it (as described in Section 2.3).

2.3. Procedure

The experiment comprised two shadowing experiments, one for each song, which were carried out consecutively.

In the first experiment, the participants were first asked to sing the Yakinton's melody with the syllable [na] instead of its lyrics. Beside that, no specific instructions were given, e.g., regarding the tempo, the key, or any other musical preference. Subsequently, the participants listened to the pre-recorded version of the song (see Section 2.1) via wired over-ear headphones (Phillips SHL3060). Following that, they sung the song once more and answered some questions regarding the recorded version of the song, to determine how much it differs from the one in their mental memory. Importantly, no reference to either their previous production or the recorded version was made by using wordings like "repeat", "mimic", "like before", etc.

The second experiment comprised only a shadowing performance, as the participants were intentionally unfamiliar with the neutral lullaby. After listening to the pre-recorded version of the lullaby, they were instructed to sing it themselves. This required not only their singing capabilities, but also their musical memory. As explained in Section 2.1, this song was specifically composed with common characteristics of the genre in mind and should therefore contain similar melodic and harmonic contents to the song in the first experiment.

3. Analysis

Since the participants aimed to produce specific musical notes (as opposed to non-specific absolute frequency while speaking), tones were used for measuring pitch instead of raw Hz values. To increase the pitch resolution, quarter tones (QTs) were used
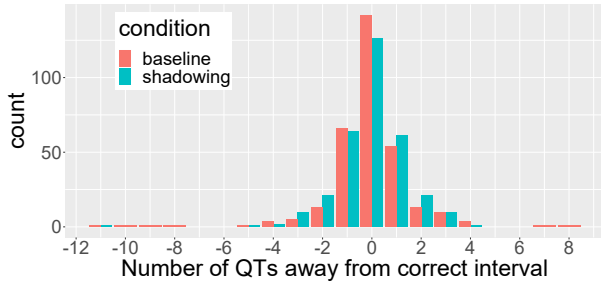
Figure 1: *Comparison between the distribution of deviations from the correct intervals in baseline (red) and shadowing (blue) performances. The numbers on the x-axis are the number of QTs above or below the correct interval.*

Figure 2: *Comparison between the deviation distribution of each interval in baseline (top) and shadowing (bottom) conditions. The numbers on the x-axis are the interval indices representing the 53 intervals in the Yakinton lullaby. The distances between the intervals sang by the participants and the correct intervals are shown on the y-axis (outliers are omitted). The labels "A" to "C" mark the different parts of the song (and correspond to the same labels in Snippet 1).*

instead of semitones. This enables a more fine-grade analysis, which can capture more subtle off-key singing to differentiate the performances better. Furthermore, segmentation of the performance into individual tones was done manually. Silences, non-singing, breaths between phrases, etc. were segmented as well. The tone frequencies were determined by the median of the measured frequencies during this tone's duration, excluding the first and last 10 % of the tone duration. This excludes transitions between tones and smooths out vibrato and *ad lib* ornaments. These values were extracted using Praat [24] and manually verified. Subsequently, the note assigned to each singing segment was determined by selecting the closest QT to the measured frequency in the corresponding segment. The mapping between tone frequencies and QTs (relative to middle A) was done using the formula

$$frequency(QT_n) = 440 \cdot \sqrt[24]{2}^n, \tag{1}$$

where *n* is the number of QTs away from the middle A key (cf. De Klerk [25]) and 440 Hz is the frequency of middle A. This formula is based on the equal temperament. QTs are denoted here with the symbols ♮♩ and ♯♩ for one QT and three QTs above a note, respectively. Ultimately, tonal deviations were measured per interval, rather than per tone, as the latter would depend on the key the participants chose, while the former measures tonal accuracy independently of key.

Tempo was measured for an entire performance, taking into account only singing segments (similar to articulation rate in speech). This ensures that pauses between phrases do not influence the perceived singing tempo, and that occasional, non-scripted lengthenings (e.g., short ritardandos and fermate at the end of phrases) do not mark a specific note as being out of rhythm. Tempo was measured in beats per minute (BPM), which is directly derived from the standard musical notation ♩=, using the formula

$$BPM = \frac{N + \delta}{overall\ duration} \cdot 60, \tag{2}$$

where *N* is the number of beats in the song and *δ* is the number of beats added in a specific performance. Such additions occurred exclusively at the end of phrases (bars 3, 6, 10, and 13 in Snippet 1), but are not present in the pre-recorded stimulus. The Yakinton lullaby (Snippet 1) and the neutral lullaby (Snippet 2) have 26 quarter beats and 32 dotted quarter beats, respectively.
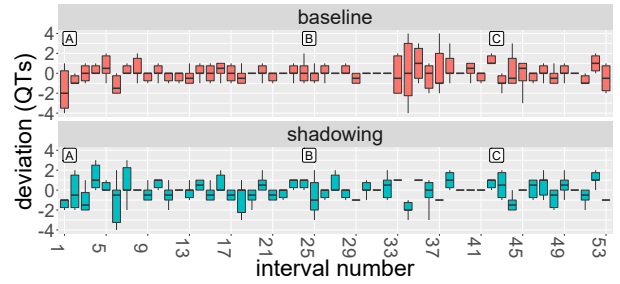
4. Results

As expected, the participants could, for the most part, accurately produce the Yakinton lullaby (Snippet 1) based merely on their mental representation of it in the baseline phase. However, as Figure 1 shows, these performances included several large deviations of two tones or more, which are not likely to be caused by coincidental imprecise singing. In the shadowing condition, in comparison, there was only one such large deviation. This adjustment of obviously wrong tones was apparently driven by the exposure to a correctly sung version. Other than these corrections, the deviation distributions shown in Figure 1 are roughly symmetric and equal to baseline in shadowing conditions. Surprisingly, the baseline performances had more tones within ±1 QT (which for the participants' western ear would still sound correct). It seems, therefore, that the reference version helped the participants to sing within a more accurate range of tones, but also to show more variation within this range.

The tone-by-tone comparison presented in Figure 2 sheds more light on these differences. It is evident that except for the very first interval, the participants showed greater consecutive variation in the second part of the bridge (label "B" in Snippet 1, notes 34 to 42), while in the shadowing condition the first phrase (first seven intervals) show a similar tendency. Although the bridge moves to a new tonal center, it is not clear why only the second part of it would cause the singers to be less precise. As for the higher variation at the beginning of the shadowing performances, this points to the process of re-finding the right tones in the participants' key of preference. This explanation is supported by the key comparisons in Table 1, which show that there was virtually no key change between the baseline and shadowing performances by any of the participants. Despite that, the unstable beginning of the shadowing performances indicates that listening to the recorded version influenced the participants' tonal accuracy. They needed about one whole phrase to overcome the influence of the different tonal center in the recorded version and go back to their key of preference. Interestingly, the only participant who sang in the same key as the recording did change key in the second performance.

The tempo of the recorded version was 61 BPM. In the baseline performance, three participants sang faster than that and three more slowly (see Table 1). All participants changed their tempo so that it was closer to the recorded version. More-

Table 1: *Comparison between the singing tempo and key in baseline and shadowing performances of each participant. The values on the left and right under the key and BPM columns are for baseline and shadowing performances, respectively. BPMΔ shows the BPM difference between baseline and shadowing, with the value in parentheses standing for the change in the difference from the recording's tempo. A negative value means that the participant decreased the distance to the recording.*

Participant	Key	BPM	BPMΔ
RITRAF85	F ∣ F♯	76 ∣ 70	6 (−6)
TALHAR82	B ∣ B♭	57 ∣ 63	6 (−2)
RANVI88	A ∣ A	59 ∣ 63	4 (0)
ONKASH82	F♯ ∣ F♯	76 ∣ 69	7 (−7)
LIIT82	F♯ ∣ F♯	59 ∣ 66	7 (+3)
DIHAR83	F♯ ∣ F♯	62 ∣ 61	1 (−1)
recording	B ∣ (B)	61 ∣ (61)	

Table 2: *Comparison between the percentage of occurrences of each rhythmic pattern in the original and replicated versions in all bar-level patterns. Parts A and B refer to the labels with the same letters in Snippet 2. Each replication row refers to the average over all participants who replicated that part.*

	R1	R2	R3	R4
original part A	50	12.5	12.5	25
replications A	54	18	7	21
original part B	25	37.5	12.5	25
replications B	25	42	8	25

over, the absolute distance from the recording's tempo decreased in all cases but one.

In contrast to the first song, it was not expected that participants would be able to completely replicate all rhythmic elements in the second song. Two participants managed to replicate part A (cf. Snippet 2), part B was replicated by three participants, and one participant replicated both part A and B. The replication rate of each rhythmic element was measured separately instead of the overall tempo. Table 2 summarizes the occurrences of each rhythmic pattern R1–R4 (corresponding to the rhythmic patterns in bars 1, 9, 15, and 16 in Snippet 2, respectively) in the original and replicated versions. The proportion of each pattern within a part was almost completely preserved in the replication, with the exception of a small difference between R2 and R3 in part A. It is also evident that R1 and R4 were replicated more accurately. An explanation for that is that they appear at the beginning and end of the song, making them easier to remember. They are also technically easier to produce, since they contain fewer tones (and intervals) than R2 and R3.

The accuracy of tonal replication was measured in two ways, viz. directionality and quantity. First, the correctness of the contour direction in each interval was evaluated (higher tone, lower tone, or same tone). Second, the size of each interval was compared with the correct interval. The participants correctly produced the contour direction in 70 % of the intervals they replicated. The intervals themselves, however, were correct only in 44 % of the times. This suggests that overall contours of the song are more easily recalled than the specific intervals. Snippet 3 shows a few specific examples of tonal and rhythmic deviations.

5. Discussion and conclusion

The experimental results presented in this paper show that convergence occurs in singing, more so with respect to temporal features than to tonal ones. This stands in contrast to findings in interactive speech [e.g., 26]. Even so, the results emphasize the similarity between the two oral capabilities, viz. speech and singing, which are both used in human communication. They are therefore prone to influence each other and can potentially be related and enhance one another. For example, Nardo and Reiterer [27, p. 216] explain that *tonal* and *rhythmic* abilities are measures of musicality and also related to phonetic tal-

ent. This idea is also supported by Tsang *et al.* [28], who found a correlation between musical experience and sensitivity to convergence. Similarly, pitch has been found to correlate with the level of agreement between interlocutors in dyadic conversation [29]. We therefore suggest that speech and music are two domains where certain common effects, and in particular convergence, occur with respect to shared prosodic features. Establishing this connection between music and speech offers a wide variety of further interdisciplinary experimentation that combine linguistic and musical analysis methods. Specifically for convergence, the influence of listening to people with different social and vocal characteristics can be examined. The manner in which distances in vocal behavior decrease, or increase, can depend on further aspects of the social environment and auditory context, as suggested by Noy [30].

In conclusion, the first hypothesis – convergence to perceived musical stimuli – was satisfied, but not in the same way for pitch and tempo. While tempo became globally closer to the recorded version in absolute terms, the tones were produced more precisely but within the same tonal range. Additionally, the secondary expectation was only partially met, with fewer large deviations occurring in the shadowing performances, but otherwise the tones in the baseline production were slightly more accurate as a whole. Finally, the third hypothesis was fulfilled, as the simpler, frequent rhythmic patterns were replicated more correctly. Furthermore, with one exception, participants were not able to replicate the entire song.

Snippet 3: *Example of tonal (top staff) and rhythmic (bottom staff) deviations in the second lullaby. Smaller, stemless notes mark the correct notes where deviation occurred. Crossed-head notes mark those that deviate from the correct rhythmic pattern.*

6. Acknowledgments

7. References

[1] J. Molino, *Toward an Evolutionary Theory of Music and Language*. The MIT Press, 2000.

[2] R. Jackendoff, "Parallels and nonparallels between language and music," *Music Perception: An Interdisciplinary Journal*, vol. 26, no. 3, pp. 195–204, 2009. DOI: 10.1525/mp.2009.26.3.195.

[3] J. Day-O'Connell, "Speech, song, and the minor third: An acoustic study of the stylized interjection," *Music Perception: An Interdisciplinary Journal*, vol. 30, no. 5, pp. 441–462, 2013. DOI: 10.1525/mp.2013.30.5.441.

[4] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006. DOI: 10.1121/1.2178720.

[5] M. Kim, W. S. Horton, and A. R. Bradlow, "Phonetic convergence in spontaneous conversations as a function of interlocutor language distance," *Laboratory Phonology*, vol. 2, no. 1, pp. 125–156, 2011. DOI: 10.1515/labphon.2011.004.

[6] L. B. Meyer, *Emotion and Meaning in Music*. University of Chicago Press, 2008.

[7] A. Schweitzer and N. Lewandowski, "Convergence of articulation rate in spontaneous speech," in *Proceedings of Interspeech 2013 (Lyon, France)*, 2013, pp. 525–529. DOI: 10.1.1.394.4806.

[8] J. S. Pardo, R. Gibbons, A. Suppes, and R. M. Krauss, "Phonetic convergence in college roommates," *Journal of Phonetics*, vol. 40, no. 1, pp. 190–197, 2012. DOI: 10.1016/j.wocn.2011.10.001.

[9] M. Babel and D. Bulatov, "The role of fundamental frequency in phonetic accommodation," *Language and Speech*, vol. 55, no. 2, pp. 231–248, 2012. DOI: 10.1177/0023830911417695.

[10] B. Collins, "Convergence of fundamental frequencies in conversation: If it happens, does it matter?" In *Fifth International Conference on Spoken Language Processing (Sydney, Australia)*, 1998.

[11] M. D'Imperio, R. Cavone, and C. Petrone, "Phonetic and phonological imitation of intonation in two varieties of italian," *Frontiers in Psychology*, vol. 5, p. 1226, 2014. DOI: 10.3389/fpsyg.2014.01226.

[12] M. Simonet, "Intonational convergence in language contact: Utterance-final F0 contours in Catalan–Spanish early bilinguals," *Journal of the International Phonetic Association*, vol. 41, no. 2, pp. 157–184, 2011. DOI: 10.1017/S0025100311000120.

[13] J. Krivokapic, "Rhythm and convergence between speakers of American and Indian English," *Laboratory Phonology*, vol. 4, no. 1, pp. 39–65, 2013. DOI: 10.1515/lp-2013-0003.

[14] S. D. Goldinger, "Echoes of echoes? An episodic theory of lexical access," *Psychological Review*, vol. 105, no. 2, p. 251, 1998. DOI: 10.1037/0033-295x.105.2.251.

[15] M. W. Weiss, S. E. Trehub, and E. G. Schellenberg, "Something in the way she sings: Enhanced memory for vocal melodies," *Psychological Science*, vol. 23, no. 10, pp. 1074–1078, 2012. DOI: 10.1177/0956797612442552.

[16] S. E. Trehub and A. M. Unyk, "Music prototypes in developmental perspective.," *Psychomusicology: A Journal of Research in Music Cognition*, vol. 10f, no. 2, p. 73, 1991. DOI: 10.1037/h0094140.

[17] J. S. Pardo, A. Urmanche, S. Wilman, J. Wiener, N. Mason, K. Francis, and M. Ward, "A comparison of phonetic convergence in conversational interaction and speech shadowing," *Journal of Phonetics*, vol. 69, pp. 1–11, 2018. DOI: 10.1016/j.wocn.2018.04.001.

[18] M. Twig, "Music as a universal language: Examining the effect of culture on music perception at infancy," Master's thesis, University of Haifa, 2016. [Online]. Available: https://haifa.userservices.exlibrisgroup.com/view/delivery/972HAI_MAIN/12137798060002791.

[19] A. M. Unyk, S. E. Trehub, L. J. Trainor, and E. G. Schellenberg, "Lullabies and simplicity: A cross-cultural perspective," *Psychology of Music*, vol. 20, no. 1, pp. 15–28, 1992. DOI: 10.1177/0305735692201002.

[20] S. E. Trehub, A. M. Unyk, and L. J. Trainor, "Maternal singing in cross-cultural perspective," *Infant behavior and development*, vol. 16, no. 3, pp. 285–295, 1993. DOI: 10.1016/0163-6383(93)80036-8.

[21] A. Fernald, "Prosody in speech to children: Prelinguistic and linguistic functions," *Annals of child development*, vol. 8, pp. 43–80, 1991.

[22] A. Street, S. Young, J. Tafuri, and B. Ilari, "Mothers' attitudes to singing to their infants," in *Proceedings of the 5th Triennial ESCOM Conference*, 2003, pp. 628–631. [Online]. Available: https://www.epos.uni-osnabrueck.de/books/k/klww003/pdfs/099_Street_Proc.pdf.

[23] M. Papoušek, H. Papoušek, and D. Symmes, "The meanings of melodies in motherese in tone and stress languages," *Infant behavior and development*, vol. 14, no. 4, pp. 415–440, 1991. DOI: https://doi.org/10.1016/0163-6383(91)90031-M.

[24] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001, version 6.1.06.

[25] D. De Klerk, "Equal temperament," *Acta musicologica*, vol. 51, no. Fasc. 1, pp. 140–150, 1979. DOI: 10.2307/932181.

[26] E. Raveh, I. Siegert, I. Steiner, I. Gessinger, and B. Möbius, "Three's a crowd? Effects of a second human on vocal accommodation with a voice assistant," in *Proceedings of Interspeech 2019 (Graz, Austria)*, Sep. 2019, pp. 4005–4009. DOI: 10.21437/Interspeech.2019-1825.

[27] D. Nardo and S. Reiterer, "Musicality and phonetic language aptitude," in *Language Talent and Brain Activity*, G. Dogil and S. Reiterer, Eds., Berlin: Mouton de Gruyter Berlin, 2009, pp. 213–256. DOI: 10.1515/9783110215496.

[28] G. J. Y. Tsang, E. L. Dana, M. M. Farbood, and S. V. Levi, "Musical training and the perception of phonetic detail in a shadowing task," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1922–1922, 2018. DOI: 10.1121/1.5036272.

[29] B. M. Okada, L. Lachs, and B. Boone, "Interpreting tone of voice: Musical pitch relationships convey agreement in dyadic conversation," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, EL208–EL214, 2012. DOI: 10.1121/1.4742316.

[30] P. Noy, *The Psychoanalysis of Art and Creativity*. Modan, 1999, in Hebrew.