SUBJECTIVE EVALUATION OF VOICE CONVERSION PERFORMANCE: TOWARDS AN UNDERSTANDING OF HOW NON-LINGUISTIC CONDITIONS CHANGE OUR VOICE

Vered Silber-Varod[1], Viktor Iaroshenko[2] and Oliver Jokisch[2]

[1]The Research Center for Innovation in Learning Technologies, The Open University of Israel
[2]Institut für Kommunikationstechnik, Hochschule für Telekommunikation Leipzig (HfTL)
vereds@openu.ac.il

Abstract: The main objective of the current study is to learn how one's voice changes according to non-linguistic conditions. This report on work in progress, aims to study the common and different qualities in non-linguistic variables by state-of-the-art Voice Conversion (VC) methods. As far as we know, this is the first report of VC tests on Hebrew speech. We first report on some baseline results of gender conversions, which were evaluated perceptually by thirty-one Hebrew native speakers. None of the original audio excerpts was evaluated correctly by 100% of the participants. In the second listening test, subjects were asked to indicate how close the voice conversion of a mimic voice (Source voice) is to an original famous figure's voice (Target voice). Results suggest that the VC process managed to shift the subjects' votes' weight from "absolutely Source voice" to relatively higher uncertain votes and higher percentage of "absolutely Target voice", compared to the votes on the original mimic voice.

1 Introduction

The main objective of the current study is to learn how one's voice changes according to non-linguistic conditions. Such extra-linguistic cues are said to reflect more fixed speaker characteristics. Exploring the varied abilities of the human voice can contribute to the understanding of the nature of verbal and nonverbal signs and symbols, and the role of communicative behavior in a variety of social contexts (including personal and organizational relationships, intercultural encounters, political and legal debates, and more).

1.1 Scientific background

Speech is the main channel by which most people communicate in their daily lives. One of the major speech communication faculties is to be able to talk differently in varied social contexts. As human listeners, we bring many sources of information to the interpretation of an utterance, including grammar, prosody, our knowledge of the world and the conversational context.

In speech sciences it is common knowledge that not only the textual content of the message matters – the what, but also the how, the voice quality a person uses in order to convey the message plays a role. In other words, the essence of social communication relies not only on communicating information or a message but also on paralinguistic and extra-linguistic information. Paralinguistic cues in the voice convey rich dynamic information about a speaker´s intentions and emotional state, and extra-linguistic cues are said to "reflect more stable speaker characteristics including identity, biological sex and social gender, socioeconomic or regional background, and age" [1].

The main objective of this proposal is to learn how one's voice changes according to sociolinguistic and non-linguistic conditions. Such understanding of the varied abilities of the human voice is part of a growing discipline called Speech communication. Speech

communication scholars strive to explore the nature of verbal and nonverbal signs and symbols, and the role of communicative behavior in a variety of social contexts (including personal and organizational relationships, intercultural encounters, political and legal debates, and more).

Most academic training in the area of speech communication takes the perspective that successful communication is partly a learned skill. Most people are born with the physical abilities to acquire necessary communication tools but such potential does not guarantee that they will learn to communicate effectively. Language, rhetorical strategies, listening skills, and a lexicon of verbal and nonverbal meanings are developed in various ways. It is theorized that people gain their communication skills by having them modeled by persons in their environment, by being taught specific techniques through the educational process, and by practicing their abilities and having them evaluated [2]. Within this framework, studies about mimicry can provide enhanced characteristics of a figure's voice and thus enhanced acoustic features that will be easier to obtain and analyzed. From a phonetic point of view, "it may tell us a lot about the flexibility of the human voice – to what extent, and in what ways, is it possible to modulate one's voice." [3].

1.2 Voice conversion technology

The main research objective of the present research is the speech signal. The speech signal carries information about two levels of the speech sound: the segmental level and the supra-segmental level. While the segmental level refers to phonemes, i.e., phones, and their incorporation into words and utterances in a given language, the supra-segmental sound patterns are "a vocal effect which extends over more than one sound segment in an utterance, such as a pitch, stress or juncture pattern" [4]. The supra-segmental sound patterns in speech are termed: speech prosody (For speech prosody research in Hebrew, see [5, 6]. Speaking of the functions of prosody, one has to take into account that speech is characterized by the co-occurrence of various features, such as fundamental frequency (F0), intensity, duration, voice quality, etc. Hence, acoustic modeling of the voice signal is one of the complex tasks in all speech technologies.

The motivation of studying the common and different qualities in human voice by Voice Conversion (hereafter, VC) methods relies first on the relatively small amount of training data to provide subjective reasonable results; Second, VC tools provide immediate audible output. Other methods of studying voice differences, such as machine learning, need huge amount of data and they provide numeral results as output (for example, the WEKA toolkit [7]). Third, voice conversion can be applied for varied purposes for the benefit of individual welfare: in the entertainment sector (e.g., personalized avatars in games); in the audio books sector, where one can create a very personal environment by modifying original voice files and emulating the sound of a familiar voice (even years after s/he has passed away), and, more important, for medical purposes [8], and in the learning sector, where it is possible to train and evaluate speech communication skills.

The goal of voice conversion (also termed speaker transformation) is to transform a sentence said by a source speaker or voice to sound as if a target speaker (or voice) had said it.

The voice conversion system gathers information from the source and target voices and automatically formulates voice conversion rules in the training stage. For this purpose, training databases from source and target voices are acoustically analyzed and a mapping between the acoustic spaces of the two voices is estimated. The transformation stage employs the mapping obtained in the training stage to modify the source voice signal in order to match the characteristics of the target voice. The modification is performed using a set of signal processing algorithms that modify the vocal tract and the prosodic characteristics [9, 10].

Other conventional VC methods are based on spectral modifications only [11, 12]. In [13], a learning-based mapping of prosodic features is suggested. It is important to note that VC methods have not yet been reported in Hebrew speech and the few VC studies in Israel are presently working on a speech database in English [10, 14]. As for other languages, VC research has already been documented in the literature in general [9, 12], and for German in particular [13, 15, 16, 17].

To sum up, the contribution of this preliminary research is to promote the methodology of VC to help individuals evaluate their speaking competence in different speaking environments. With VC methods, we can understand better the minor yet significant changes that a speaker applies in different social and cultural interfaces. By providing a tool that will process a source voice by taking into account a target voice, we can provide speech technology to one of the major learning skills – speaking.

2 Method and Research Design

2.1 Database

2.1.1 Gender test recordings

The gender conversion test consists of twelve Hebrew sentences recorded by three men and three woman. Each uttered two different sentences, one was left as is (original audio) and one was manipulated for its gender charachtersitics. The sentences vary in their length and consist of both monosyllabic utterances and syntactically coherent sentences. For example, Man3 uttered single words: [bo'i] "come" and [kfar] "villlage", while Woman1 utterd utternces [hayeled paxad lalexet levad baxoshex] " The boy was afraid to walk alone in the dark" and [hu ra'ad mikor bli me'il] " He was shivering with cold without a coat" (number of syllables per utterance are represented by the transcribed vowels, as each vowel represents a syllable).

The recordings were carried during 2013. All recordings were carried out in a quiet room, using an Audio Technica AT892 head mounted microphone in order to maintain a relatively constant mouth to microphone distance. Digitization was performed with a Centrance Micport Pro audio interface to a personal computer, maintaining constant gain levels throughout all the recordings. Sampling frequency was 44.1 kHz throughout.

2.1.2 Mimicry recordings

The mimicry test consists of 18 Hebrew sentences spoken by two speakers: 1. Binyamin Netanyahu, the prime minister of Israel at the time of recordings; 2. Tuvya Tsafir, a famous mimicry actor in Israel. Netanyahu's voice was defined as the *target* voice, and thus only three utterances were extracted from his speech. Netanyahu's voice was *not* manipulated. On the other hand, 15 utterances were extracted from the actor's speech, and his voice was defined as the *source* voice.

Both recordings were extracted from online free YouTube videos. The real Benjamin Netanyahu (BN) video was published in May 11, 2014 [18]. The video is part of a very popular satiric TV show where Mr. Netanyahu was the host of the show, and participated in a humoristic episode. The imitation of BN by the actor was published in Apr 29, 2012 [19], at that time BN was already the PM of Israel. The signals were transformed into WAV format (e. g. PCM 16 kHz, 16 bit). The provided signal quality (signal-to-noise ratio) of given video/audio sources is usually limited (compared to high quality/HQ recordings) but sufficient for the majority of VC methods (VTLN, HMM methods etc.). Since the two recordings were not taken at the same conditions, a white noise was added to mask the source (mimicry) voice, so it will be more similar to the audio quality of Netanyahu's TV show.

Two utterances from each speaker were used to train the listeners' perception (see Appendix I). The 15 source's utterances were either embedded as is, without manipulations, or were converted. The list of utterances is presented in Appendix I.

2.2 Voice conversion tool

For both VC tests PRAAT software version 6.0.05 [20] was used.

2.2.1 Gender voice conversion design and manipulation

For the gender test, two conversions types were carried out using PRAAT "change gender" tool: 1. Female to Male VC, and 2. Male to Female VC.

The conversion was carried out on the same linguistic content of the six original voices. For both Male-to-female and Female-to-male conversions, the same pitch range was used: 75 Hz as a pitch floor and 600 Hz as a pitch ceiling. The arguments that controled the manipulation were:

- Male to Female: Formant shift ratio: 1.1, new pitch median (Hz): 200, pitch range factor: 1, duration factor: 1.
- Female to male: Formant shift ratio: 1, new pitch median (Hz): 120, pitch range factor: 1, duration factor: 1.

2.2.2 Mimicry voice conversion design and manipulation

The voice conversion method in the mimicry test was carried according to the process design in PRAAT [20] as illustrated in Figure 1. The challenge in the present VC process was that the two speakers, actor (source) and famous figure (target) did not produce the same utterances.

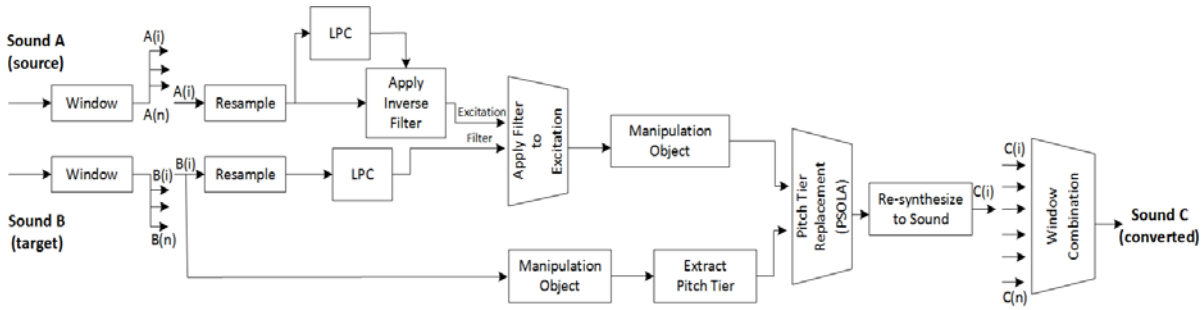

Figure 1 – Voice conversion algorithm in PRAAT [20], retrieved from: http://archive.cnx.org/contents/5a06554b-0b6e-41a3-89d1-fc0cf06a2e75@4/voice-conversion-in-praat

2.3 Listening tests

The two test forms were created using Google Form platform. The gender form design followed the test design of [21] and of [16]. Subjects were asked to make a gender decision for each sample they heard, i.e., to which extent they perceived a male (1) or a female (-1) voice. The gender decision reflected subject's opinion on the gender of the speaker in the original audio file or in the converted output. Subjects have simply rated each recording on a five-step scale according to speaker's gender: 1 – Male; 0.5 – Closer to Male; 0 – Cannot decide; -0.5 – Closer to Female; -1 – Female.

The mimicry form design followed [22]. Subjects were asked to make a speaker decision for each sample along with source and target speaker samples they heard, i.e., to which extent they perceived the Source voice or the Target voice. The speaker decision reflected subject's

opinion on the identity of the speaker in the audio sample. Subjects have simply decided whether the item was uttered by the “Source” speaker or by the “Target” speaker. Subjects rated each recording on a five-step scale according to speaker's identity: Source voice; Closer to Source voice; Cannot decide; Closer to Target voice; Target voice.

All subjects in both tests were fluent Hebrew speakers.

3 Results

3.1 Subjective identification of gender voice conversion

We first report on some baseline results of gender conversions, which were evaluated perceptually by Hebrew native speakers. Thirty-one participants answered the first subjective test on gender conversion, which consisted on twelve audio excerpts – six original males' and females' voices and six converted male-to-female and female-to-male samples. None of the original audio excerpts was evaluated correctly by 100% of the participants. The subjects' votes are shown in Figure 2.

For original male voice, the best average score was 0.968 (on a scale from -1 to 1, where -1 is "absolutely female" and 1 is "absolutely male"), and for female – 0.984 (in absolute values). Minimum average scores of original voice were 0.290 for male's voice and 0.839 for female voice. Male-to-female conversions did better than female-to-male (on average, 0.446 and 0.048, respectively).

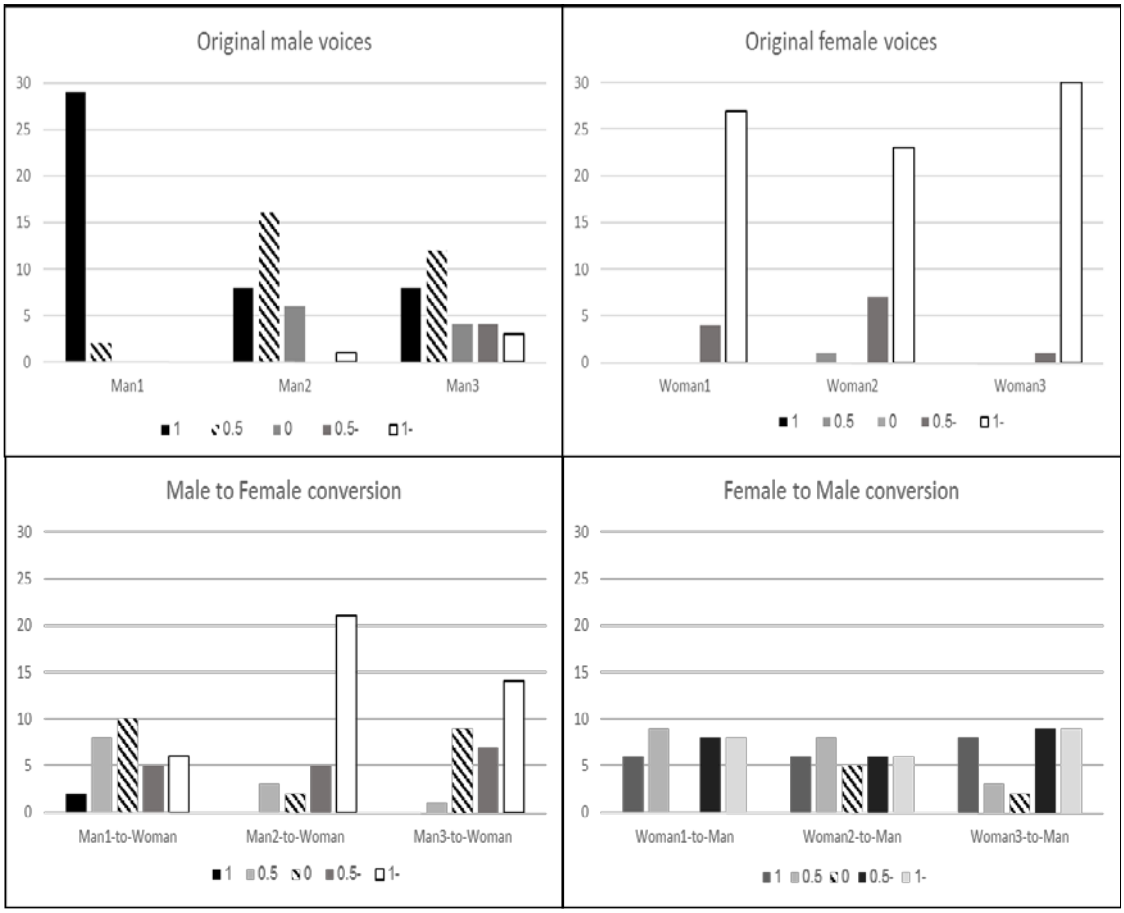

Figure 2 – Gender identification tests: Top left: Subject's votes for original male voices; Top right: Votes for original female voices; Bottom left: Male-to-Female conversions; Bottom right: Female-to-Male conversions.

3.2 Subjective identification of mimicry voice conversion

At a second stage, we conducted a subjective test of VC performance within the framework of mimicry. Thirty participants answered the second subjective test on mimic voice conversion, which consisted on 20 audio excerpts – three original Target voice of BN:

[Sirtey burekas] translated "Bourekas films"; [ata yodea xagiga basnuker] translated "You know Party at the Snooker"; [ma ata rotse xinux o otsar] translated "What do you want – Education or Treasury (ministry)?".

Four original *mimic* of BN voice. For example:

[ma yesh lexol hasmolanim haele] translated "What all these (political) lefties want?"

13 converted source-to-target (i.e., mimic-to-BN) samples. For example:

[hakol hitxil be'elef tsha meot arba'im veshmone] translated "Everything began in 1948".

As with the gender tests, none of the original audio excerpts was evaluated correctly by 100 % of the participants. The subjects' votes are shown in Figure 3. For original mimic voice, the best average score was -0.733 (On a scale from -1 to 1, where -1 is "absolutely mimic voice" and 1 is "absolutely BN voice"), and for target BN voice – 0.683. Minimum average scores of original voices were -0.217 for mimic's voice (in Test 8 [kshedaxafti et eh] "When I pushed uh…") and 0.033 for BN voice (in Test 18 [ma ata rotse xinux o otsar] "What do you want – Education or Treasury (ministry)?"). Mimic-to-BN conversions did on average -0.173.

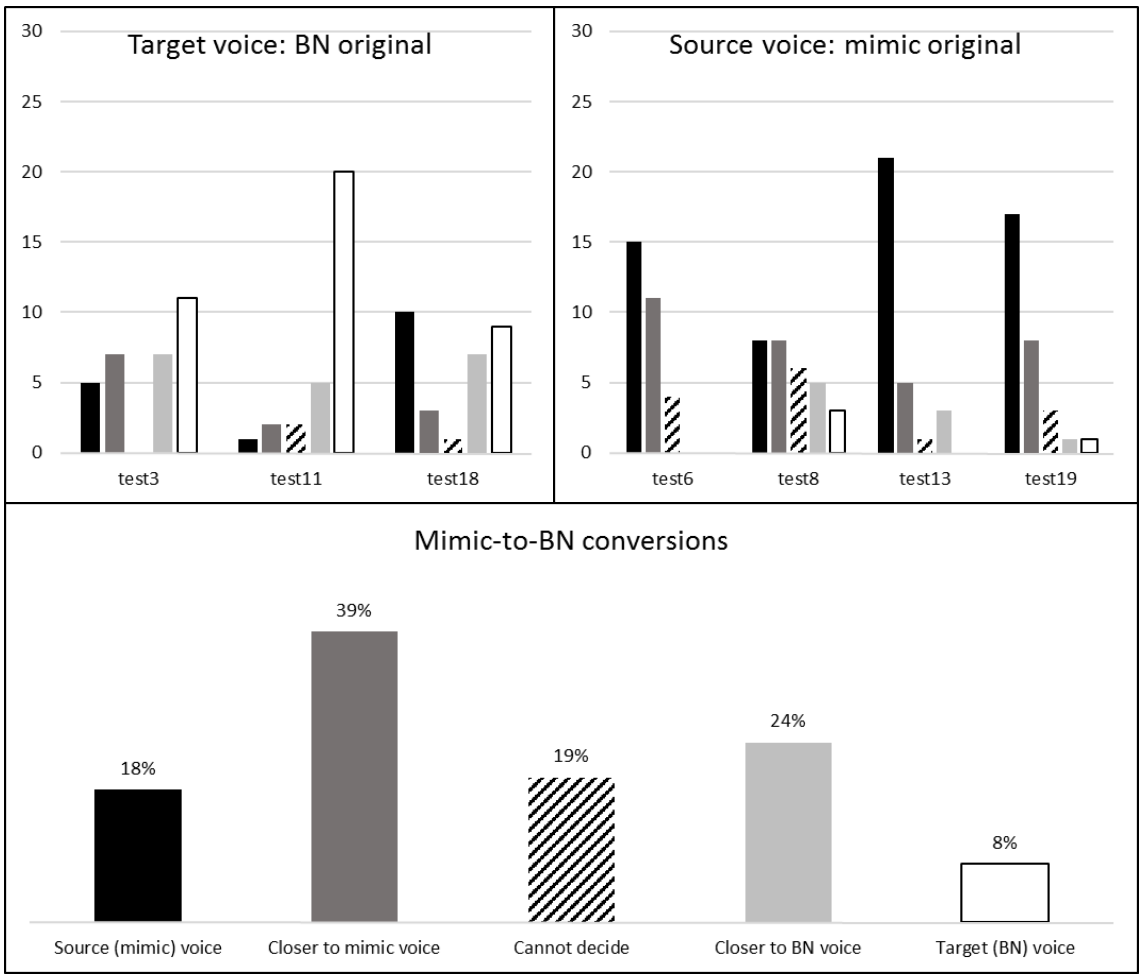

Figure 3 – Mimic to target voice identification tests: Top left: Subjects' votes for original BN voice (Target voice); Top right: Originl mimic voice (Source voice); Bottom: Total votes (%) of mimic-to-target conversion tests.

Table 1 presents a comparison between the original mimic identification votes and the mimic VC output. A shift from 55% to 18% of "absolutely Source voice" votes is evident. Moreover, more uncertain votes are evident in the VC output. Last, there is a rise in the percentage of "absolutely Target voice" votes for the VC output.

Vote scale	Original mimic	Mimic-to-Target VC
Source (mimic) voice	55%	18%
Closer to mimic voice	29%	39%
Cannot decide	13%	19%
Closer to BN voice	8%	24%
Target (BN) voice	4%	8%

Table 1 – A comparison of the original mimic votes to the VC votes

4 Discussion

In this preliminary VC research on voice identification judgements of Hebrew spoken samples, we showed that listeners do not identify speakers' gender nor speaker's identity in full. The baseline gender VC test showed that male-to-female conversions did better than female-to-male's. Since we used the same VC method, this can be explained by the lower rates of original males' identification compared to original females' identification rates (i.e., identification of two male voices turned out to be a challenge to the listeners, and as a consequence their conversion to female voices yielded high votes).

In the source-to-target conversion test, correct identification of both *original* voices, mimicry of BN and BN himself, suggest that the high quality of the mimicry in actor's voice confused the subjects as only the train samples got relatively high rates (64% for BN voice in Test 11 and 71% for mimic voice in Test 13). Another possible reason for this confusion is the context of the sentences. BN sentences were uttered during his participation in a humoristic TV show and thus do not reflect the usual political agenda context.

Subjects' votes on the mimic-to-BN conversion samples demonstrate a shift from relatively high percent of "absolutely mimic voice" to more uncertain votes, and to higher percentage of "absolutely Target voice", as demonstrated in Table 1. Thus, the VC process did manage to shift the votes' weight.

A byproduct of the present research is the understanding that VC technology, and speech technologies in general, can be used as a research tool and not only as an end product. With VC as a research tool we can learn about the perceptual aspect of human voice and about the varied parameters that are combined into the subjects' perception process: Prosody and the amount of physical material – speech, and interestingly – the context.

In future research we intend to investigate the acoustic parameters that will yield a better identification of non-linguistic variables that undergo voice conversion process.

5 Acknowledgement

This research was funded by the German Academic Exchange Service (Deutscher Akademischer Austauschdienst – DAAD) within the program Research Stays for University Academics and Scientists, 2015 (50015559).

References

[1] SCHWEINBERGER, S. R., KAWAHARA, H., SIMPSON, A. P., SKUK, V. G., ZÄSKE, R.: *Speaker perception.* Wiley Interdisciplinary Reviews: Cognitive Science, 5(1), pp. 15 – 25, 2014.

[2] MORREALE, S. P., OSBORN, M. M., PEARSON, J. C.: *Why communication is important: A rationale for the centrality of the study of communication.* Journal of the Association for Communication Administration (JACA) 29, pp. 1 – 25, 2000.

[3] ERIKSSON, A., WRETLING, P.: *How flexible is the human voice? – A case study of mimicry.* Target, 30(43.20), pp. 29 – 90, 1997.

[4] CRYSTAL, D.: *Dictionary of Linguistics and Phonetics*, sixth edition. Blackwell 2008.

[5] SILBER-VAROD, V.: *The SpeeCHain Perspective: Form and Function of Prosodic Boundary Tones in Spontaneous Spoken Hebrew.* Lambert Academic Publishing, 2013.

[6] SILBER-VAROD, V.: *Structural analysis of prosodic pattern: The case of excessive prolongations in Israeli Hebrew.* Revista Leitura, Spec. Issue on Speech Prosody, Vol. 52, 271 – 291,2013. http://www.seer.ufal.br/index.php/revistaleitura/article/view/1483/1011

[7] HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., WITTEN, I. H.: *The WEKA data mining software: an update.* ACM SIGKDD Explorations Newsletter, 11(1), pp. 10 – 18, 2009.

[8] NAKAMURA, K., TODA, T., SARUWATARI, H., SHIKANO, K.: *Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech.* Speech Communication, 54(1), pp. 134 – 146, 2012.

[9] TÜRK, O.: *Cross-lingual voice conversion.* PhD thesis, Bogaziçi University, 2007.

[10] BENISTY, H., MALAH, D., CRAMMER, K.: *Modular global variance enhancement for voice conversion systems.* Proc. 20th European Signal Processing Conference (EUSIPCO), pp. 370 – 374, Bucharest, 2012.

[11] PITZ, M., NEY, H.: *Vocal Tract Normalization Equals Linear Transformation in Cepstral Space.* IEEE Transactions on Speech and Audio Processing 13, pp. 930 – 944, 2005.

[12] STRECHA, G., JOKISCH, O., EICHNER, M., HOFFMANN, R.: *Codec integrated voice conversion for embedded speech synthesis.* Proc. INTERSPEECH, 2589 – 2592, 2005.

[13] SCHWARZ, J.: *Statistische Stimmenumwandlung in Kombination mit prosodischen Modellen.* PhD thesis, Christian-Albrechts-Universität Kiel, 2010 (in German).

[14] BENISTY, H.: *Voice Conversion with Enhanced Temporal Spectrum-Variability.* PhD thesis, Technion – Israel Institute of Technology, 2015. http://webee.technion.ac.il/people/DavidMalah/Pubs/Grad/Hadas_Benisty/Statistical_Voice_Conversion_Short_Summary.pdf

[15] JOKISCH, O., GAMBOA, H.: *Performanzuntersuchungen zur Stimmkonvertierung.* Proc. ESSV Conference (Studientexte zur Sprachkommunikation 61), Aachen, pp. 349–356, 2011 (in German).

[16] JOKISCH, O., BIRHANU, Y., HOFFMANN, R.: *Runtime and speech quality survey of a voice conversion method.* Proc. IEEE Conf. EUROCON, pp. 1690 – 1694, Zagreb, 2013.

[17] SÜNDERMANN, D., STRECHA, G., BONAFONTE, A., HÖGE, H., NEY, H.: *Evaluation of VTLN-based voice conversion for embedded speech synthesis.* Proc. INTERSPEECH, Lisbon, pp. 2593 – 2596, 2005.

[18] YOUTUBE *Binyamin Netanyahu:* http://youtu.be/JLj3TSao_YE

[19] YOUTUBE *Tuvya Tsafir:* http://youtu.be/jIe6dVxrcjQ

[20] BOERSMA, P., WEENINK, D.: *Praat: doing phonetics by computer* [Computer program]. Version 6.0.05, retrieved 4 November 2015 from http://www.praat.org/

[21] TÜRK, O., ARSLAN, L. M.: *Subjective evaluations for perception of speaker identity through acoustic feature transplantations.* Proc. EUROSPEECH, pp. 2093 – 2096, 2003.

[22] HUBER, S., RÖBEL, A.: *Voice quality transformation using an extended source-filter speech model.* Proc. 12th Sound and Music Computing Conf. (SMC), pp. 69-76, 2015. Voice conversion listening test on French: http://stefan.huber.rocks/phd/tests/VoCoX2F/