The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 296-309

Influences of Prosodic Feature Replacement on the Perceived Singing Voice Identity

Kuan-Yi Kang
Department of Electrical Engineering
National Tsing Hua University
sjk102061231@gapp.nthu.edu.tw

Yi-Wen Liu
Department of Electrical Engineering
National Tsing Hua University
ywliu@ee.nthu.edu.tw

Hsin-Min Wang
Institute of Information Science
Academia Sinica
whm@iis.sinica.edu.tw

Abstract

Human perception on the singing voice differs with the factors of the singing voice and the subjects. On one hand, the background knowledge influences the understanding of voice for each subject. On the other hand, the difference of the voices presented to the subjects also affects the perception. In this paper, we discuss two factors reflecting on the similarity before and after singing voice conversion: prosodic features and subjects' familiarity to the singers. Three experiments were conducted. The first experiment tested the subjects' ability to identify the singer. The second experiment synthesized the singing voice with different singers' prosodic features, and let the subjects score the similarity. The third experiment presented timbre-converted singing voice with different combinations of prosodic features from two singers to the subjects for them to judge the similarity to the target singer.

The results show that, first, the number of prosodic features contained in the synthesized voice is positively correlated with the scores in identification and similarity. Also, subjects who are more familiar personally with the target singers have better identification scores than target-unfamiliar subjects on the timbre-converted singing voices.

Keywords: Singing voice conversion, Prosody, Human perception, Voice identity.

1. Introduction

In the task of voice conversion, subjective tests examining quality and similarity [1] are used to evaluate the synthesized results from human perception. The quality test asks the subjects to score how great the quality of the converted result is, while the similarity test questions about how similar the result is, comparing to the target speaker.

The acoustic characteristics of voice individuality can be described through timbre, pitch, intensity, duration, etc. [3]. Hence, the synthesis of the voice considers not only the timbre but also other prosodic features such as pitch, duration or intensity [4]. What effects the modification of these features have on the identification of speakers [5] and how these features represent individuality [6,7,8] have also been discussed in previous works.

Studies in voice perception have investigated on the acoustic features affecting voice identity with experiments using the correlation analysis and multidimensional scaling techniques [12]. The modifications of acoustic features [13,16] and vocal identity aftereffects [14] are also used to determine the relative importance factors on humans' ability of speaker identification.

While some studies on the contributing features of voice identity contain timbre and other prosodic features [12,13,16], this paper investigates only on the prosodic features in order to find out the essential prosodic parameters for changing the perceived singer identity. In addition, since that the majority of the voice conversion tasks only focus on the development of the timbre conversion [1], using the source prosodic features for the synthesized voice, we therefore want to examine how changing the prosodic features might potentially help convert the voice more convincingly, and how that would reflect on the subjective similarity test.

The result of a subjective test depends on the background knowledge of the participants. When the subjects know more about the audio presented, the knowledge might influence the test performance; in [2,15], the difference of the human ability on identifying familiar and unfamiliar voices was well discussed. Therefore, we would also like to examine how familiarity to the speaker reflects on the similarity performance in tasks related to the singing voice conversion.

In this paper, we follow a similar experiment design of perception test in [11] by mixing up the features of source, target or converted singing voice to discuss the effects of sound

modification on the perception. Three listening tests for the participants are designed in order to find out the effects due to the subjects' familiarity to the singers and due to the changes of prosodic features. Section 2 describes the experiments designed. Section 3 discusses the listening test results and the effects of different factors. Section 4 then gives the conclusion.

2. Methods

2.1 Recordings

The recordings consist of parallel singing voice data sung by two female singers, F1 and F5. The recording process was conducted in a quiet room with a microphone, an RME interface and Cubase software. Each singer sang 9 pop songs, a total length of 17 minutes. These recordings were further tuned using Cubase and cut into phrases, each phrase ranging from 7 to 15 seconds.

2.2 Participants

We invited 17 participants to the listening test using their own laptops and headphones. 53% had a music training background (have learned musical instruments for more than one year) and 47% did not. 35% of participants were familiar with the voices of the two singers before the listening test, 24% knew one of them, and 41% of subjects were unfamiliar with the voices of the two singers. 47% of subjects were familiar with singer F1, and 47% of subjects were familiar with F5.

2.3 Experiment design

Before starting the test, all the participants were presented with a one-minute singing clip from both singers in order to be acquainted with both singers' singing voices. These audio files could be replayed during the test if the participants wanted to re-learn the singer's voice.

2.3.1 Singer identification

In the first task, we aimed to examine the participants' ability to distinguish signers from their singing voice. The participants were presented with 6 phrases from each singer in random order. The participants would then be asked to distinguish whether the audio presented was sung by F1 or F5.

2.3.2 Identification and similarity task of the timbre-carrying singer

The second task was designed to examine how the prosodic and timbre features would influence human perception, and how the changes of prosodic features would affect the identification task. We fixed the timbre of one singer and selectively replaced the expressive features (pitch, intensity, and duration) from the singer to the other singer. Out of the three expressive features, there were 8 kinds of combinations for feature replacement. For each combination, there were 4 examples (2 examples for a singer). The feature replacement combinations and their nomenclature are summarized in Table 1. Singer A indicates the original timbre-carrying singer, and singer B is the singer whose features are used to substitute singer A's features. In total, 32 audio files were thus prepared for participants to listen to.

The fundamental frequency and spectral envelope of a singing voice were extracted with the WORLD vocoder [10]. We used the fundamental frequency as the pitch feature. The spectral envelope was further compressed into mel-cepstral coefficients, with the first dimension defined as the intensity feature and the other dimensions defined as the timbre feature in this experiment. The duration features were modified through the dynamic time wrapping (DTW) conducted on the timbre feature if the selected identity was singer B; the pitch and intensity of the target would be adjusted to the source length through DTW if the duration identity was singer A. The selected and modified features were then synthesized into the singing voice with the vocoder.

Table 1. Feature Combination of Synthesized Voice

Nomenclature	**Timbre**	**Pitch**	**Intensity**	**Duration**
AAAA	Singer A	Singer A	Singer A	Singer A
AAAB	Singer A	Singer A	Singer A	Singer B
AABA	Singer A	Singer A	Singer B	Singer A
AABB	Singer A	Singer A	Singer B	Singer B
ABAA	Singer A	Singer B	Singer A	Singer A
ABAB	Singer A	Singer B	Singer A	Singer B
ABBA	Singer A	Singer B	Singer B	Singer A
ABBB	Singer A	Singer B	Singer B	Singer B

For each audio file that was listened to, the participants were asked to perform singer identification and score the similarity. The identification task asked which singer sang the audio, in the format of ABX test, while the voice of the two singers were introduced in the

section before the first experiment (2.3.1). The similarity test asked how similar the audio was to the timbre-carrying singer with the minimum opinion score (MOS).

2.3.3 Identification and similarity task of the timbre-converted singer

Existing voice conversion algorithms mostly apply original dynamic changes of source prosodic features on the synthesis results with the timbre converted through models [1]. The third task was to examine how the conversion of the expressive features may play some roles in the similarity test. Based on the timbre converted from the source to the target using a Gaussian mixture model [9], the experiment here tested 8 kinds of feature combinations on pitch, intensity, and duration, shown in Table 2, to test the effects of the prosodic features on human perception. The source singer is F5 and the target singer is F1. Each type contains 3 different singing phrases with timbre converted, so 24 files in total were presented to the participants.

The usage of each expressive feature follows a similar procedure in Sec. 2.3.2. The frame-based spectral features of the source singer were converted with a Gaussian mixture model [9]. The expressive features would be used directly without modification if the features' assigned identity was the source. If the target's duration feature was used, all the other features would be adjusted to match the target's length based on the DTW alignment of the spectral features of the two singers. The target's pitch and intensity could be extracted with the vocoder, and the length would be adjusted through DTW if the duration scale was assigned as the source.

Table 2. Origin of Features of Synthesized Voice

Nomenclature	**Timbre**	**Pitch**	**Intensity**	**Duration**
CSSS	Converted	Source	Source	Source
CSST	Converted	Source	Source	Target
CSTS	Converted	Source	Target	Source
CSTT	Converted	Source	Target	Target
CTSS	Converted	Target	Source	Source
CTST	Converted	Target	Source	Target
CTTS	Converted	Target	Target	Source
CTTT	Converted	Target	Target	Target

For each audio file they listened to, the participants were also asked two questions,

identification and similarity, with the identification task forcing the listener to determine which singer produced the audio, and the similarity task asking the listener to score how similar the audio was to the target singer.

3. Results of experiments

3.1 Singer identification

The performance of the subjects on the identification task of the original singers is shown in Table 3. The overall accuracy of all subjects is 80.39%. When the subjects are more familiar with the singers before the experiment, they will perform better on the identification task. The subjects who knew both singers achieved 95.83% accuracy, while the subjects not knowing any of the singers had 64.29% accuracy. The subjects with a music background were 8% more accurate than the subjects without a music background.

Table 3. Identification accuracy % (mean ± std)

All subjects	
All	80.39±20.61
Familiarity to the singers	
Knowing 0 singer	64.29±20.81
Knowing 1 singer	85.42±14.23
Knowing 2 singers	95.83±6.97
Music background	
Without music background	76.04±23.33
With music background	84.26±18.37

3.2 Identification and similarity task of the timbre-carrying singer

The modification of the expressive features from the original singer could change the perceived singer identity. Using 8 combinations of expressive features, without the conversion of timbre in the singing voice, the scores for each type (4 examples) were first averaged for each subject. Then, for each type, the average scores of the 17 subjects were analyzed to obtain the mean and standard deviation. The identification and similarity scores for each combination of expressive features are shown in Figure 1.

As can be seen from Figure 1, when some prosodic features are replaced, the identification score and the similarity score do decrease, especially AABB, ABAA, and

ABBB have the lowest scores compared with AAAA. The reason why the scores of AAAB and ABAB are comparable to the scores of AAAA could be that the two singers sang in a similar style in the audio files randomly selected for these types, so that after replacing one or two expressive features, the clip could still be recognized as the original singer.

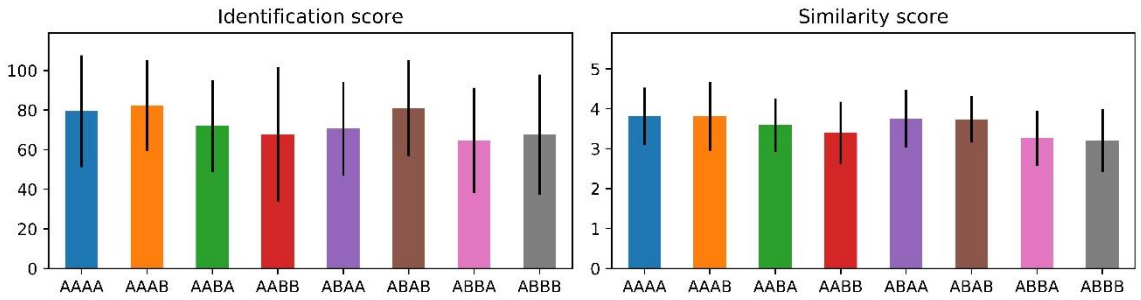

Figure 1. The identification score (%) and similarity score (mos) of the timbre-carrying singer with changed prosodic features.

The dependency of the scores upon the number of replaced expressive features was analyzed and shown in Table 4. Category 0 denotes AAAA while category 3 (all 3 expressive features were replaced) denotes ABBB. Category 1 (1 expressive feature was replaced) consists of AAAB, AABA, and ABAA, while category 2 (2 expressive features were replaced) consists of AABB, ABAB, and ABBA. For categories 2 and 3, the mean and standard deviation of 17 subjects and 3 feature combinations were calculated. It is clear that the identification and similarity scores decrease as more changes were made to the original singer's expressive features. When the number of replaced features increases from 0 to 3, the identification score drops from 79.41% to 67.65% and the similarity score drops from 3.81 to 3.21.

Table 4. Identification and similarity scores (mean ± std) of the timbre-carrying singer with different numbers of replaced expressive features

Number of replaced features	Identification Score %	Similarity Score
0	79.41±28.23	3.81±0.72
1	75.00±23.45	3.72±0.75
2	71.08±28.88	3.47±0.71
3	67.65±30.32	3.21±0.78

Table 5 shows the results with a specific expressive feature replaced. There are 3

expressive features: pitch, intensity and duration. For each feature, the mean and standard deviation of 68 samples (17 subjects and 4 feature combinations) were calculated. For example, for types AABA, AABB, ABBA, and ABBB, the "intensity" was replaced. Compared to duration, changes in pitch and intensity have lower scores and greater reduction from AAAA, indicating a greater impact on human perception of singer individuality.

Table 5. Identification and similarity scores (mean ± std) of the timbre-carrying singer with a specific expressive feature replaced

Features changed	Identification Score %	Similarity Score
Pitch	70.96±26.39	3.49±0.73
Intensity	68.01±28.27	3.36±0.73
Duration	74.63±28.49	3.54±0.78

Since each subject had different levels of prior knowledge about the singing voices that were presented, we divided the subjects into 3 categories: familiar with none, 1, or 2 of the singers before the experiments. The score distributions are depicted in Figure 2. Over all, the scores increase when the number of familiar singers increases. In other words, for the cases where timbre was unchanged but some expressive features were changed, the subjects familiar with more singers had better performance on singer recognition. Even with the changes in expressive features, the subjects who know the two singers have an identification score of higher than 70%, suggesting that the subjects tend to rely on the timbre as the cue to identify the singer.

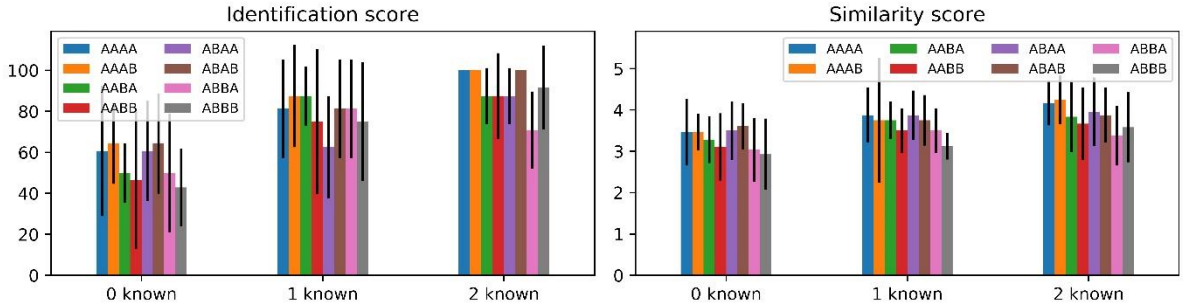

Figure 2. The identification score % and similarity score of the timbre-carrying singer with changed prosodic features and the number of familiar singers.

3.3.3 Identification and similarity task of the timbre-converted singer

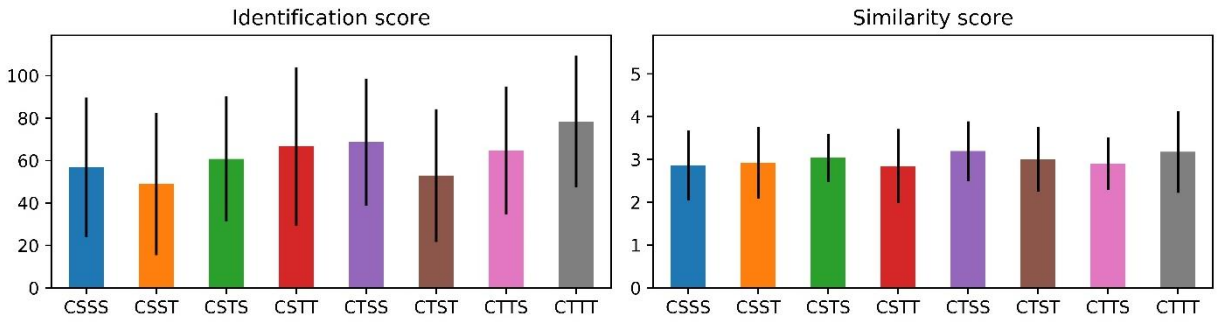

Figure 3. The identification score % and similarity score of the timbre-converted singer with changed prosodic features.

In the conversion task, the expressive features used in the synthesized results could lead to different perceptions in different subjects. For the 8 expressive feature combinations with converted timbre, the scores of each type for each subject were first averaged. We then calculated the statistics of the scores among all 17 subjects for each type. The results are shown in Figure 3.

The similarity score is around 3.0, meaning that the subjects were neutral on the decision of the similarity to the target singer, although the score seemed to slightly increase when more target expressive features were used. For the identification score, the situation CTTT achieved the best performance, which was 78.41% and was 21.55% higher than CSSS as shown in Table 6, suggesting that using all the three kinds of target prosodic features led to much better identification scores than using only the source prosodic features.

Table 6. Identification and similarity scores (mean ± std) of the timbre-converted singer using different numbers of target expressive features

Number of changed features	Identification Score %	Similarity Score
0	56.86±32.84	2.86±0.82
1	59.48±31.49	3.05±0.70
2	61.44±32.91	2.92±0.74
3	78.41±31.05	3.18±0.96

The scores with different numbers of target expressive features used are shown in Table 6. The calculation of the results was the same as in Table 4. It is clear that the identification score increases when more target expressive features are used. The score slightly increases

when using only 1 or 2 target expressive features, and the usage of all the target prosodic features improves the score from 56.86% to 78.41%.

Table 7 shows the results when a specific type of target expressive feature was used in the converted singing voice. The synthesized singing voice including target's pitch and intensity has higher probabilities to be identified as target than the synthesized singing voice including target's duration. The results reconfirm that pitch and intensity have a greater impact of human perception of singer individuality than duration.

Table 7. Identification and similarity scores (mean ± std) of the timbre-converted singer using a specific target expressive feature

Feature changed	Identification Score %	Similarity Score
Pitch	66.18±31.28	3.07±0.76
Intensity	67.65±32.05	2.99±0.76
Duration	61.76±34.68	2.99±0.85

Table 8. The identification score % of the timbre-converted singer with changed prosodic features and subjects' familiarity to the singers. (Subjects: 9 unfamiliar, 8 familiar)

	Target-unfamiliar	Target-familiar
CSSS	55.56±37.27	**58.33**±29.55
CSST	51.85±29.40	45.83±39.59
CSTS	62.96±20.03	58.33±38.83
CSTT	51.85±33.79	**83.33**±35.63
CTSS	70.37±20.03	66.67±39.84
CTST	44.44±28.87	**62.50**±33.03
CTTS	62.96±26.06	**66.67**±35.63
CTTT	77.78±28.87	**79.17**±35.26
All	59.72±29.04	**65.10**±35.85

The effects of subjects' familiarity to the singers on different feature combinations are shown in Table 8. While considering only the factor of target-familiarity (averaging the scores over subjects and feature combinations), the identification score of the target-familiar subjects is 5.38% higher than that of the target-unfamiliar subjects. The target-familiar subjects achieved higher identification scores than the target-unfamiliar subjects in most

combination types. For the cases of CTSS, CTTS and CTTT, the differences between two groups are relatively small and both have the accuracy higher than 60%.

3.3.4 Discussion

In the second experiment (cf. Sec. 3.3.2), the change of the prosodic features in the singing voice of the original timbre-carrying singer degraded the performance of the identification task. In the third experiment (cf. Sec. 3.3.3), the type CTTT achieved the best performance among the 8 types. When more expressive features of the target were used in the converted singing voice, the identification score for the target singer improved. Both experiments show that the modification of pitch and intensity have higher influences on human perception of singer individuality than duration.

Subjects' familiarity to the singing voice also influences the identification task. In the first experiment (cf. Sec. 3.3.1), the subjects who were familiar with the singers personally achieved higher identification scores than the subjects who were not familiar with the singers. In the second experiment, even with the changes in expressive features, the singer-familiar subjects seemed to be able to rely on the timbre for identification, thus maintaining an identification score of greater than 70% (for the subjects who know both singers). The third experiment showed that the target-familiar subjects had a higher identification score than the target-unfamiliar subjects and more subjects in target-familiar groups successfully recognized the synthesized results as produced by the target singer when converted timbre and target expressive features were used.

Each feature combination type in the experiment consisted of 3 or 4 randomly selected files from the data set, and the scores discussed were based on the answer of the 3 or 4 files. For some phrases, the two singers might sing in a similar way with less individuality in the singing voice. If these kinds of files were selected, the identification would depend more on the individuality of timbre itself rather than the expressive features. In the future, more audio files of each types and more subjects could be included in order to cover more situations so we may have stronger conclusions supported by rigorous statistical analysis.

4. Conclusion

In this paper, three perception experiments were designed to find out the influence of the expressive features and the familiarity of the subjects to the singer on the perceived singer

identity in the synthesized singing voice. Identification and similarity tests were conducted.

We found that, first, with the timbre unchanged, the modification of the expressive features to another singer degraded the performance on the identification task. Secondly, during the voice conversion task, the identification scores for the target singer improved when more expressive features of the target were used in the converted singing voice. In addition, in the task of examining the timbre-converted voices, the target-familiar subjects had a higher identification score than the target-unfamiliar subjects when target expressive features were used.

From these experiments, we therefore conclude that not only the timbre but also the expressive features play a role on capturing the singer's identity in voice perception, and the subjects' familiarity to the voice presented also influences the results. The task of voice conversion should also take the conversion of expression features and the subjects' familiarity to the voice into consideration in the future development. In addition, to support our findings by rigorous statistics, more coverage of audio files shall be used and more subjects can be included in future work.

5. Acknowledgements

We would like to thank Prof. Shan-Hung Wu of National Tsing Hua University for the support on the singing voice research. We are also grateful to Hsin-Te Hwang and Yu-Huai Peng who provided suggestions on voice conversion that greatly assisted the research.

References

[1] S. Mohammadi, A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, no. Supplement C, pp. 65–82, 2017.

[2] S. Schweinberger, et al. "Speaker perception," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 5, no. 1, pp. 15–25, 2014.

[3] H. Kuwabara and Y. Sagisak, "Acoustic characteristics of speaker individuality: Control and conversion," *Speech Communication*, vol. 16, no. 2, pp. 165–173, 1995.

[4] M. Schröder, "Emotional speech synthesis: A review," *Proc. Eurospeech*, pp. 561–564,

2001.

[5] Y. Lavner, I. Gath, and J. Rosenhouse, “The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels,” *Speech Communication*, vol. 30, no. 1, pp. 9–26, 2000.

[6] L. He and V. Dellwo, “Between-speaker variability in temporal organizations of intensity contours,” *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 488–494, 2017.

[7] V. Dellwo, A. Leemann, and M.-J. Kolly, “Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors,” *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1513–1528, 2015.

[8] A. Leemann, M.-J. Kolly, and V. Dellwo, “Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison,” *Forensic Science International*, vol. 238, pp. 59–67, 2014.

[9] T. Toda, A. W. Black, and K. Tokuda, “Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[10] M. Morise, F. Yokomori, and K. Ozawa, “World: A vocoder based high-quality speech synthesis system for real-time applications,” *IEICE Transactions on Information and Systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[11] Z. M. Smith, B. Delgutte, and A. J. Oxenham, “Chimaeric sounds reveal dichotomies in auditory perception,” *Nature*, vol. 416, no. 6876, pp. 87–pp.70, 2002.

[12] O. Baumann and P. Belin, “Perceptual scaling of voice identity: common dimensions for different vowels and speakers,” *Psychological Research PRPF*, vol. 74, no. 1, pp. 110, 2010.

[13] G. Sell, C. Sujed, M. Elhilali, and S. Shamma, “Perceptual susceptibility to acoustic manipulations in speaker discrimination,” *The Journal of the Acoustical Society of America*, vol. 137, no. 2, pp. 911–pp.922, 2015.

[14] M. Latinus and P. Belin, “Perceptual auditory aftereffects on voice identity using brief vowel stimuli,” *PLoS One*, vol. 7, no. 7, pp. e41384, 2012.

[15] S. R. Mathias and K. von Kriegstein, “How do we recognise who is speaking?,” *Front Biosci (Schol Ed)*, vol. 6, pp. 92–109, 2014.

[16] Y. Lavner, I. Gath, and J. Rosenhouse, “The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels,” *Speech Communication*, vol. 30, no. 1, pp. 9–26, 2000.