Interspeech 2024
1-5 September 2024, Kos, Greece

Familiar and Unfamiliar Speaker Identification in Speech and Singing

Katelyn Taylor[1], Amelia Gully[1], Helena Daffern[2]

[1]Department of Language and Linguistic Science, University of York, York, UK
[2]AudioLab, School of Physics, Engineering and Technology, University of York, York, UK
katie@ntaylor.co.uk, amelia.gully@york.ac.uk, helena.daffern@york.ac.uk

Abstract

Little research has been conducted to gauge a listener's ability to recognise or identify speakers when presented with samples of singing within the field of Forensic Speech Science. Eight friends and two foil speakers were recorded speaking and singing to investigate the effects of speaker familiarity and singing in speaker identification tasks. The stimuli were used to create a listening test completed by close social network speakers, members of the wider social network, and general lay listeners. The study aimed to explore the impact of familiarity on an individual's ability to recognise speakers when presented with spoken and sung stimuli. The results revealed that the listeners within the close social network were the most successful in the listening test. Overall, listeners performed best when both samples were spoken, however, those in the close social network were less affected by the use of sung samples, and scored higher, compared to those outside the close social network.

Index Terms: speaker identification, speaker familiarity, cross-modal identification, forensic speech science, singing

1. Introduction

Understanding the impact of speaker familiarity in speaker recognition and identification, as well as the use of samples of singing, is relatively under researched within the context of Forensic Speech Science. Speaker identification tasks test whether a witness can identify the voice of the perpetrator from the scene of a crime (Brown 1979; Yarmey, 1995). Erickson (2016), Foulkes and Barron (2000), Peynircioğlu et al. (2017) and Doromal (2016) have conducted research into the role of speaker identification within a close social network (CSN), as well as the effect of using singing and cross modal speech samples in a speaker identification task. However, there is little research that has explored the role of singing alongside speaker familiarity. Furthermore, this gap in literature and previous research leads to an uncertainty about whether an individual's ability to successfully recognise or identify a speaker is more influenced by their familiarity with the speaker, or the mode of speech they are presented with during the listening test.

The purpose of a *speaker identification* task is to test whether the witness can identify the voice of the suspect, and to verify the validity of the witness' memory whilst *speaker recognition* refers to a listener's ability to recognise a speaker due to a general familiarity with one or more of the voices within the sample of speech (Brown, 1979; Yarmey, 1995). Although voice parades may not occur as often as the standard identity parades which ask witnesses to identify a person's face, the identification or lack thereof can be paramount in determining the outcome of a case where the suspect's voice is key to identifying them as guilty (Smith et al., 2020).

Whilst there has been some research into the ability to discriminate between voices when both samples are spoken speech in a variety of conditions, there is relatively little around speaker recognition when listeners are presented with something other than spoken speech, in particular - singing. Peynircioğlu et al. (2017) set out to test whether the speaking voices of unfamiliar people could be matched to their singing voices. Overall, the results showed that the success rate was lower when one sample was spoken and the other was sung, as opposed to being the same mode where either both samples were spoken or both sung. Likewise, the findings of Doromal (2016) found that changing the mode of speech from the perceived "normal" will negatively impact the participants' ability to recognise speakers.

Meanwhile, Foulkes and Barron (2000) assessed individuals' abilities to identify speakers where all the participants were friends. The overall success rate for the listeners was 68%, equating to 61 out of 90 correct answers. There was an error rate of 10%, and no individual listener managed to identify all the speakers correctly, the best performer correctly identifying nine voices, but falsely identified the 10th member of the network as himself. The results found that whilst some voices were easily identified by all the listeners, others were more difficult, with one speaker not being able to identify themself. Similarly, Erickson (2016) conducted a study to determine the effect that familiarity of a speaker had on a listener's ability to identify whether a singer was unknown. The results found that completing familiarisation training prior to the test did not significantly improve the ability to distinguish between voices when they were compared to other singers with similar timbres. However, when the voices differed in timbre and characteristics, training did significantly improve the listener's ability to discriminate between voices.

By exploring the role of both speaker familiarity and the use of singing within speaker identification, it is possible to test whether the listener's familiarity of the speaker or the mode of speech presented to the listener have a greater impact on their ability to successfully identify the speaker. Furthermore, the results will indicate whether there is an impact from the differing modes of speech across all three social network categories.

2. Methodology

2.1 Speakers

The audio stimuli were produced by 10 native British English students, aged between 19 and 33. All the speakers were assigned female at birth; however, one speaker identifies as male. The speakers all have accents typical to the UK, with seven of the 10 speakers claiming to have a Southern Standard British English accent or an accent from the south of England, and the author verified that no accents stood out in a pilot study. The speakers can be separated into two categories - speakers

10.21437/Interspeech.2024-1763

within the CSN, and speakers acting as foils who were outside of the social network. The eight speakers within the CSN are all part of the University of York Central Hall Musical Society (CHMS) and are often in rehearsals with each other for up to 20 hours per week, in addition to any further contact they have as part of their degrees or through general socialisation. Meanwhile, the two foils are students on the MSc in Forensic Phonetics course who have never met the other speakers. The foils are familiar with each other's speaking voices, due to time spent together in class and socially, however, they had limited exposure to each other's singing voices. To provide an approximate measure of vocal equivalence, each speaker performed a vocal range test for their singing. An analysis of their spontaneous speech was also carried out in Praat (Boersma & Weenink, 2022) to calculate their average fundamental frequency (F0).

Table 1: A list of the sung vocal range and fundamental frequencies (F0) of the speakers.

Speaker	F0 (Hz)	Vocal Range
1	196	E3 – A5
2	156	D3 – C6
3	198	A3 – E5
4	188	C3 – D6
5	186	D3 – C6
6	201	D3 – D6
7	167	E3 – F5
8	189	G3 – A5
9	188	E3 – B5
10	195	E3 – G6

2.2 Stimuli

As part of the stimuli collection process, the speakers were asked to tell the story of Cinderella to elicit spontaneous speech with relatively controlled content, as the speakers were not told what they would be asked to say in advance. For the singing, each participant was then asked to sing three sections of "Over the Rainbow" from *The Wizard of Oz* (Lloyd Weber et al., 2011) followed by two sections of "Borrowed Time" from *Death Note: The Musical* (Wildhorn & Murphy, 2023). All stimuli were recorded in one recording studio using the same equipment, which included a DPA headset microphone, a -10dB attenuator, and a Zoom F8N interface. Participants were played the starting note(s) before each take which were all sung acapella.

2.3 Listening Test

The listening test consisted of two sets of data, spoken speech and singing. Four samples of speech, each three seconds, were selected from the retelling of Cinderella which allowed the participants in the listening test to have a close comparison between the speakers given the similar content in each of the samples. For the samples of singing, three samples were chosen due to the similar notes and style of the extract, with *Over the Rainbow* alternating between an E4 and a G4 and *Borrowed Time* alternating between a D4 and an E4. After each of the audio samples had been cut to three seconds using Audacity (Audacity Team, 2023), a short 5ms fade-in and fade-out was added to each recording. All samples were also normalised to -20dB RMS level to ensure none of the recordings stood out in the listening test due to their perceived amplitude.

Participants completed an online listening test using Qualtrics (Qualtrics XM, 2023). Both the speakers and lay listeners completed the same test. The listening test contained 40 questions and an additional two familiarisation questions. In each question, the speakers were given a reference audio sample (the "reference sample speaker", RSS), and then asked to choose which of 10 audio samples was the same individual (the "disputed speaker samples", DSS). Participants could listen to the samples as many times as they wished. Four conditions were tested within the survey: speaking-speaking (SpSp), where both RSS and DSS were recordings of spontaneous speech; Speaking-singing (SpSi), where the RSS was spontaneous speech, but the DSS was a sample of singing; singing-speaking (SiSp), where the RSS was a sample of singing and the DSS was spoken speech; and finally singing-singing (SiSi), where both the RSS and DSS were recordings of singing. Each condition had 10 questions, one with each speaker as reference, for a total of 40 questions. When analysing the results, participants were awarded scores. If they chose the correct answer they were given one point, otherwise no points were awarded. Each participant can therefore score a minimum of 0 and a maximum of 40.

The listening test was piloted with the two foil speakers who had recorded stimuli for the listening test but were outside of the CSN as participants. They highlighted some audio differences for certain speakers, including a poorer recording quality and amplitude. Therefore, these samples were re-cut and edited before being added to the survey.

2.4 Participants

The listening test was open for four weeks and opportunity sampling was used to recruit participants by distributing the test online and allowing anyone to participate. Participants were asked to disclose whether they were an active member of CHMS, as society membership would increase the likelihood of participants being more familiar with the speakers' singing and spoken voices, which in turn could have an impact on their score. At the close of the test, full responses were collected from 93 participants, a breakdown of which can be found in Table 2.

Table 2: A breakdown of the participant demographics from the listening test.

Categories	Number of Participants
Total Responses	93
Male	22
Female	67
Non-Binary	4
Close SN	8
Wider SN	18
Outside SN	67

3. Results

3.1 Social Network

The success rate (Figure 1) of those in the CSN was similar to the results of Foulkes and Barron (2000) and Erickson (2016) who found that familiarity with the speakers did improve the likelihood of the participant identifying the correct speaker. This would indicate that participants performed with a higher degree of success the more familiar they were with the speakers.

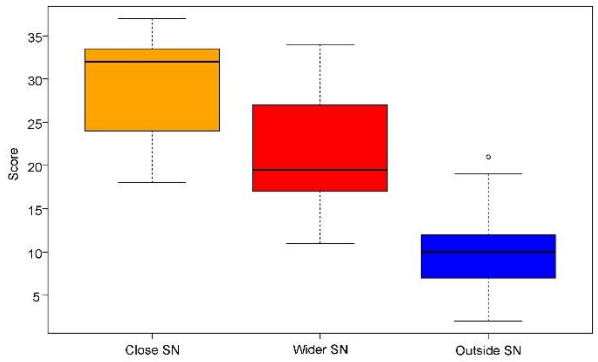

Figure 1: Box plot illustrating the difference in average score by social network (close, wider, and outside) out of a possible 40 correct answers.

A one-way ANOVA comparing the effect of the participants' social network on their overall score revealed that there was a statistically significant difference in the score between at least two social networks (F (2, 90) = [90.27], p < 2.22e-16). A Tukey HSD test found that the mean score was significantly different between the CSN and wider social network (WSN) (Adjusted p-value = [7.96e-04], 95% C.I. = [-12.26, -2.79]), CSN and outside social network (OSN) (Adjusted p-value = [4.35e-10], 95% C.I. = [-23.32, -14.98]), and WSN and OSN (Adjusted p-value = [4.35e-10], 95% C.I. = [-14.58, -8.66]). The calculated eta-squared was 0.667, which indicates that the social network condition has a large effect on the score. Overall, this signifies that as the familiarity with the speakers decreased, so did the mean score from the listening test. A linear mixed-effects model also revealed that social network was the only factor that impacted the listener's confidence in their answer during the listening test, which they indicated using a 6-point Likert scale.

3.2 Mode of Speech

As there were 40 audio excerpts in total, comprising four speaking/singing conditions, each participant was able to score a maximum of 10 for each condition and a total score out of 40. The participants were most successful in the SpSp condition, with over 57% of the total answers correct where both samples were spoken. Contrariwise, the rate of success dropped for the other three conditions, with the SiSi condition performing only slightly better (32%) than the singing/speaking conditions (24% and 25% respectively), as further demonstrated in Figure 2.

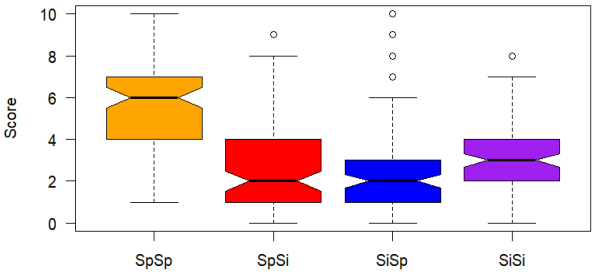

Figure 2: Box plot demonstrating the difference in mean score (out of 10) achieved by the participants in each condition (Speaking-Speaking, Speaking-Singing, Singing-Speaking, Singing-Singing).

Following on from the initial findings, a one-way ANOVA was performed to compare the effect of the singing/speaking condition on the participants' ability to select the correct answer. The ANOVA revealed that there was a statistically significant difference in the score between at least two of the conditions (F (3, 368) = [40.07], p < 2.22e-16).

The subsequent Tukey HSD test found that the mean score was significantly different between SpSp and SpSi (Adjusted p-value = [<< 0.001], 95% C.I. = [-4.04, -2.28]), SpSp and SiSp (Adjusted p-value = [<< 0.001], 95% C.I. = [-4.17, -2.41]), and SpSp and SiSi (Adjusted p-value = [9.69e-12], 95% C.I. = [-3.36, -1.59]). The eta-squared also resulted in an effect size of 0.246, signifying that the condition had a substantial effect on the score. There was no statistically significant difference in the mean scores between SpSi and SiSp (adjusted p-value = 0.98), SpSi and SiSi (adjusted p-value = 0.18), or SiSp and SiSi (adjusted p-value = 0.08).

3.3 Social Network and Mode of Speech

An analysis of the combined effect of the participants' social network and the question condition was also carried out.

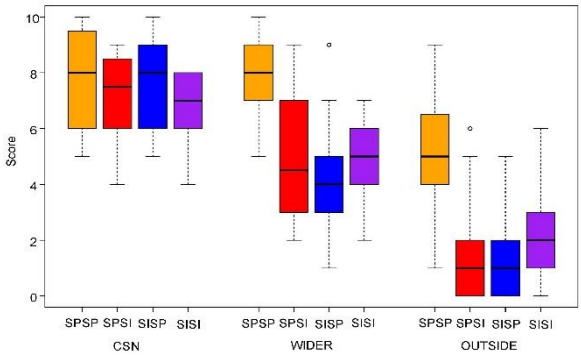

Figure 3: Box plot showing the mean scores (out of 10) per condition (SpSp, SpSi, SiSp, SiSi) within each of the social networks (close, wider, outside).

A one-way ANOVA was carried out to compare the influence of the different question conditions on the score for the CSN. The ANOVA analysis did not reveal a significant effect (F (3,28) = [0.57], p = 0.64) however, the one-way ANOVA comparing the different question conditions on the mean score for the WSN did (F (3,68) = [11.72], p = 2.79e-06). The Tukey HSD test found that the WSN mean scores were significantly different between SpSp and SpSi (Adjusted p-value = [2.47e-04], 95% C.I. = [-4.54, -1.13]), SpSp and SiSp (Adjusted p-value = [3.83e-06], 95% C.I. = [-5.26, -1.85]), and SpSp and SiSi (Adjusted p-value = [3.34e-04], 95% C.I. = [-4.48, -1.07]). There was no statistically significant difference in the WSN mean scores between SpSi and SiSp (adjusted p-value = 0.68), SpSi and SiSi (adjusted p-value = 0.99), or SiSp and SiSi (adjusted p-value = 0.63).

Another one-way ANOVA was performed to compare the effect of the four conditions on the score within the OSN. The ANOVA revealed that there was a significant difference between the question conditions and the score for at least two of the groups (F (3,264) = [82.02], p < 2.22e-16). The

subsequent Tukey HSD test found a statistically significant difference in mean score between SpSp and SpSi (Adjusted p-value = [7.57e-14], 95% C.I. = [-4.23, -2.87]), SpSp and SiSp (Adjusted p-value = [7.57e-14], 95% C.I. = [-4.28, -2.92]), SpSp and SiSi (Adjusted p-value = [1.14e-13], 95% C.I. = [-3.25, -1.89]), SiSi and SpSi (Adjusted p-value = [1.30e-03], 95% C.I. = [0.30, 1.67]), and SiSi and SiSp (Adjusted p-value = [6.82e-04], 95% C.I. = [0.35, 1.71]). There was no significant difference between the mean scores of SpSi and SiSp (adjusted p-value = 0.998).

4. Discussion

The results for the different social network categories determined that as the participants' familiarity with the speakers decreased, so did their mean speaker score. The results also indicated that there was a large effect on the mean score between all three social network categories. Further analyses suggested that the CSN were able to successfully identify many of the speakers on multiple occasions despite it being an open set, as was also found in Foulkes and Barron (2000). There were instances where the CSN participants struggled to identify certain speakers which may have been a result of similar F0s between the speakers, as well as similar features in their speech and singing, including their accent, the use of l-vocalisation or t-glottalisation, and the presence or absence of vibrato in their singing. Overall, Speaker 7 achieved the highest number of correct identifications scoring 37 out of a possible 40. Meanwhile, Speaker 9 was the most correctly identified speaker in all social network categories, with 35 correct identifications out of 40 by other members of the CSN, and an overall success rate of 56% by all participants, with 210 identifications out of 372. The results of the present study are somewhat comparable to those of Foulkes and Barron (2000) and Erickson (2016) who found that listening test participants with some degree of familiarity with the speakers had a relatively high success rate in speaker identification, and that this rate was higher than that of the OSN participants. These results cannot be directly compared to any previous studies as very little research has been conducted on the impact of familiarity using singing.

An analysis of the listening test data found the condition of the audio files presented to the participants during the test did have a substantial effect on their score. Whilst there was no significant difference between SpSi, SiSp and SiSi, the results indicated the mean score in the SpSp condition was significantly higher when compared to the other three conditions. This would suggest the participants found it easier to recognise or identify a speaker when both samples were spoken compared to when either one or both samples were sung. Overall, the results are in line with what was partly suggested by the findings of Peynircioğlu et al. (2017) which found that participants performed worse in the cross-modal conditions. Unlike Peynircioğlu et al. (2017), these results also imply that the unimodal condition where both samples were sung did not lead to a higher mean score, as they originally suggested, given the statistically insignificant differences between the SpSi, SiSp and SiSi conditions. Instead, the results indicate that where no sample of speaking is present, the success rate of the participants is the same as the cross-modal condition. This could be due to the use of different songs for the RSS and DSS, and the slight variations in musical genre between the two songs.

An analysis was carried out to determine the interaction between the social network of the participant, the speaking/singing condition, and their combined effect on the score. Whilst there was no significant difference found between the modality conditions in the CSN condition, a significant effect was found in the WSN and OSN groups. In the WSN group, a significant effect was identified between the SpSp and other three conditions, but there was no significant effect between the cross modal and SiSi conditions. Meanwhile, in the OSN group, the findings revealed a significant effect between all conditions, except for the two cross modal conditions, SpSi and SiSp. Therefore, these results imply that in the CSN, where the participants are most familiar with the speakers, the modality had no significant impact on the mean score. In the WSN, the participants all scored higher on average in the SpSp condition compared to the other three conditions, which would suggest that the cross-modal conditions did decrease the chances of the participants selecting the correct answer, as predicted by Peynircioğlu et al. (2017). Unlike their results, however, these results indicate that it was not the unimodal conditions that aided the participants in achieving a higher score, but rather the use of speech over singing.

Alternatively, in the OSN condition where there was a significant effect in the SpSp and SiSi conditions, the results once again show that participants scored higher on average in the SpSp condition. Furthermore, the significant result between SpSi and SiSi, as well as the SiSp and SiSi conditions, provides evidence to suggest that for the OSN participants, the use of unimodal samples (either both speaking or both singing) does increase the chance of them correctly matching one of the disputed sample speakers with the reference speaker.

5. Conclusions

The results of this study have several important implications for the field of Forensic Speech Science. The results indicate that within a CSN, friends can identify each other from their speech and their singing, which is partially in line with the findings of Foulkes and Barron (2000). It is important to note that this is a slightly unusual scenario where the social network knows each other because of their membership in a musical theatre society, meaning that they are also accustomed to hearing each other sing. Future research could investigate the extent to which the friends remain able to identify each other using different styles of speech, such as read or shouted; and using different styles of singing, including belted or whispered singing. More research could also be conducted to investigate whether listeners perform better when they are only familiar with a participant's speaking voice.

An analysis of the speakers' speech and singing also revealed that certain features, such as the speaker's accent or use of vibrato, may help earwitnesses to make a correct identification. Despite these features, the results still demonstrated that not all participants were able to pick-up on these, and therefore were still likely to incorrectly identify the speaker. Moreover, the variation in score between all participants within each social network condition also suggested that, despite being presented with the same material, some people are inherently better at identifying speakers compared to others. Finally, the results identified that participants in the listening test struggled to correctly identify the speaker when one sample was sung, and the other was spoken. Therefore, this would suggest that in forensic cases where the evidence includes one sample of spoken speech and one sample of singing, the earwitness is less likely to successfully identify the suspect. Instead, the present results suggest that the earwitness would have the highest chance of correctly identifying the witness when presented with two samples of speech.

6. References

[1] Brown, R. (1979). Memory and decision in speaker recognition. *International Journal for Man-Machine Studies, 11*, 729-742

[2] Yarmey, A. D. (1995). Earwitness speaker identification. *Psychology, Public Policy, and Law*, *1*(4), 792.

[3] Erickson, M. L. (2016). Can listeners hear who is singing? The role of familiarity. *Journal of Voice*, *30*(5), 638-e7.

[4] Foulkes, P., & Barron, A. (2000). Telephone speaker recognition amongst members of a close social network. *International Journal of Speech, Language, and the Law*, *7*(2), 180-198.

[5] Peynircioğlu, Z. F., Rabinovitz, B. E., & Repice, J. (2017). Matching speaking to singing voices and the influence of content. *Journal of Voice*, *31*(2), 256-e13.

[6] Doromal, M. (2016). Bilingual and whispered speaker identification within a social network. MSc dissertation, University of York.

[7] Smith, H. M., Bird, K., Roeser, J., Robson, J., Braber, N., Wright, D., & Stacey, P. C. (2020). Voice parade procedures: Optimising witness performance. *Memory*, *28*(1), 2-17.

[8] Boersma, P., & Weenink, D. (2022). *Praat: doing phonetics by computer* [Computer program]. Version 6.3.03. http://www.praat.org/

[9] Lloyd Webber, A., Arlen, H., & Stothart, H. (2011). *The Wizard of Oz*. London, United Kingdom: London Palladium.

[10] Wildhorn, F., & Murphy, J. (2023). *Death Note: The Musical* [English Production]. London, United Kingdom: London Palladium.

[11] Audacity Team. (2023). *Audacity: Free Audio Editor and Recorder* [Computer programme]. Version 3.3.3. https://audacityteam.org/

[12] Qualtrics XM. (2023). *The leading experience management software*. https://www.qualtrics.com/?rid=cookie&prevsite=uk&newsite=en&geo=IN&geomatch=au