Evaluating the effects of modified speech on perceptual speaker identification performance

Benjamin O'Brien, Christine Meunier, Alain Ghio

Evaluating the effects of modified speech on perceptual speaker identification performance

Conference Paper · September 2022

DOI: 10.21437/Interspeech.2022-463

3 authors:

Benjamin O'Brien
Université d´Avignon et des Pays du Vaucluse
27 PUBLICATIONS **31** CITATIONS
SEE PROFILE

Christine Meunier
Aix-Marseille Université
75 PUBLICATIONS **781** CITATIONS
SEE PROFILE

Alain Ghio
French National Centre for Scientific Research
154 PUBLICATIONS **824** CITATIONS
SEE PROFILE

Interspeech 2022
18-22 September 2022, Incheon, Korea

Evaluating the effects of modified speech on perceptual speaker identification performance

Benjamin O'Brien[1,2], *Christine Meunier*[1], *Alain Ghio*[1]

[1] Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
[2] Laboratoire Informatique d'Avignon, Université d'Avignon, Avignon, France

benjamin.o-brien@univ-amu.fr

Abstract

This paper details a study to evaluate the effects of modified speech on perceptual speaker identification (SID) performance by naive listeners. Speech recordings made by eight male, native-French speakers were selected from the PTSVox database. The pitch and speech tempo of the recordings were modified at the word-level. The first 75% of words spoken were modified, such that the percentage of modification began at 100% and gradually decayed to 0%. The direction of the modifications was also examined, such that pitch modifications began at $\pm$600 cents and speech tempo modifications began at a ratio of either 1:2 or 3:2 (modified to normal speech tempo). Following a familiarization period, participants completed two rounds of 48 "go/no-go" task trials (balanced), where each round corresponded to a different speech modification type. The main results showed perceptual SID performance was significantly affected when participants were presented speech recordings that contained pitch modifications in comparison to speech tempo modifications. The findings revealed participants were able to overcome higher percentages of speech tempo modifications to make correct distinctions between speakers. Although modified pitch influenced in voice perception performance, high variability between participant responses were observed, which suggests listeners model speakers differently.

Index Terms: voice perception, modified speech, digital signal processing, speaker identification

1. Introduction

Recent literature reviews on perceptual speaker identification (SID) [1][2][3][4][5] have highlighted the distinction between listeners who are familiar or unfamiliar with speaker voices. Several studies reported that despite factors, such as inter-speaker variability [6] or background noise [7][8], listeners were more consistent at discriminating familiar speakers in comparison to unfamiliar ones. In general, both listener- and speaker-dependent models have been proposed to explain this facility [3], where listeners perceive certain vocal characteristics and use them to build accessible speaker representations. Although *naive* listeners are incapable of locating "unfamiliar" speaker representations, most have the capacity to discriminate between different speakers by comparing voice qualities and perceiving some as similar as opposed to others [9][10]. However, the methods used by naive listeners to perceive and gauge voice similarities are still not very well understood.

The cues used by naive listeners to distinguish speakers are multiple, as speaker identity perception has been linked to processing various acoustic features, such as fundamental and formant frequencies [11][12][13] and phonetic content [14][15][16]. Thus, it is difficult to attribute the weight of perceptual SID skills to any one of them. Studies on speech production have shown that the fundamental frequency (F0) characterises speakers according to their sex [17], age [18] or socio-cultural background [19]. A study by [13] demonstrated that F0, F2, and F3 all play important roles in speaker identification. More recent work has revealed similar findings, as a study by [11] showed naive listeners used different perceptual similarity dimensions when identifying female (F0 and F1) and male speakers (F0 and mean difference between F4 and F5). These findings underscore the importance of pitch on perceptual SID, however, as evidenced, various dimensions and their combinations are used differently by listeners.

Similarly, listeners have been shown to be sensitive to very small variations in speech rate [20] and rhythmic variations that constitute speaker-specific indexical information [21, 22]. However, in general, speech is characterized by a very large amount of variation. If some speakers speak slower than others, the amount of variation in speech tempo between speakers can be quite large [20, 23].

Although both pitch and speech tempo are cues used by listeners to identify speakers, it is important to underline that they used by speakers for communicative purposes. Thus, their variation is not exploitable as such by listeners to distinguish between speaker voices.

The current study aimed to evaluate the relative weight of pitch and speech tempo in the recognition of learned voices. By modifying the pitch and speech tempo of speech recordings made by a set of speakers, the goal was to examine their effects on perceptual SID performance by naive listeners. While it was hypothesized that modified pitch might play a greater role in influencing perceptual SID performance, it was of interest to examine thresholds across both pitch and speech tempo dimensions. Any observations would be crucial for understanding voice perception better.

2. Methods

2.1. Speakers

Speech recordings from eight native-French speakers (all male) were selected from the PTSVox database [24]. The speaker age range was 18 to 24 years (mean age 20.5 $\pm$ 2.0 years). Speaker details are available in Table 1. All speakers recited three French-texts: "Ma soeur est venue chez moi hier", "Au nord du pays, on trouve une espèce du chat"; and "La bise et le soleil se disputaient". They were recorded with a Zoom H4N stereo microphone (sampling rate: 44.1 kHz; bit depth: 16-bit).

Four speakers were assigned to an in-set speaker group, while the remaining speakers were assigned to an out-of-set group, where in-set speaker voices were used for "go" trials

10.21437/Interspeech.2022-463

Table 1: *Speaker descriptions*

Speaker	Set	Age	Region (FR)	F0 (Hz)	Speech tempo (pho/s)	Distance (ED)
LG004	Out	22	Rhône	117	13.5	3.8 ± 2.9
LG005	In	18	Rhône	101	13.8	12.3 ± 2.8
LG008	Out	24	Lorraine	114	13.7	2.4 ± 1.6
LG010	Out	19	Donzère	111	15.3	2.6 ± 2.8
LG013	In	22	Loire	125	14.2	11.8 ± 2.9
LG016	In	20	Isère	138	14.8	24.8 ± 2.1
LG019	Out	20	Loire	111	14.4	2.5 ± 2.7
LG024	In	19	Rhône	110	12.3	4.2 ± 2.1

and out-of-set speaker voices were used for "no-go" trials (see 2.5.3). These groupings were made to create cohesion in terms of the (dis)similarities in speech characteristics between each in-set speaker and all of out-of-set speakers. To do this, calculations of each speaker's fundamental frequency (F0) and speech tempo were made. The former was determined by implementing a YIN algorithm [25] in MATLAB 2016b (MathWorks Inc, USA). To obtain the latter, the mPraat Matlab toolbox [26] was used to first load speech recordings and speech-analysis Praat [27] Textgrid files into MATLAB and then calculated speech tempo (phones per second). Finally these metrics were used to calculate the euclidean distance (ED) between speakers. The out-of-set speakers had a mean ED value of 3.78 $\pm$ 1.95 ranging from 0.9 (between LG010 and LG019) to 6.26 (between LG004 and LG010). The mean ED value between in-set and out-of-set speakers was 13.26 $\pm$ 7.99 ranging from 4.2 $\pm$ 2.1 (LG024) to 24.8 $\pm$ 2.1 (LG016).

2.2. Stimuli

For each in- and out-of-set speaker 12 speech fragments were extracted from speech recordings using Praat. Although recordings contained speech read from the same texts, each was unique (96 total), e.g. the phonetic content inherent to similar texts was always presented out of phase by a minimum of 4 words. This was done to ensure speech modifications did not bias any words (see Section 2.3 for more details). The duration of the extractions ranged from 10.9 to 18.0 s (mean duration 14.3 $\pm$ 1.4 s). All 96 speech recordings were normalised with a peak to 0 dB with MATLAB. This peak normalization was performed on each recording by adjusting the maximal amplitude to a target of 100% of the signal dynamic.

2.3. Speech modification processing

Several reasons contributed to the decision to design a method that began each recordings with 100% speech modifications and then gradually decayed to 0% - no modifications, normal speech - once 75% of the total number of words were uttered. First the method was preferred over traditional methods that, for example, might present to listeners speech recordings with totalities of 0%, 25%, 50%, 75%, and 100% speech modifications, as the latter would require us to develop more stimuli for additional trials. In turn, this might introduce memory bias or fatigue to listeners. Second the method afforded listener responses to be examined across a continuum, which could provide crucial information as to inter- and intra-listener sensitives to different speech modifications. The decision to only modify the first 75% of words in the speech recordings was to provide listeners with the remaining 25% of words un-modified (typical speech), which they could use to affirm (or reject) their responses and re-familiarize characteristics associated with each speaker. Finally, the decision to modify words instead of phones was to preserve locally prosodic information contained in each spoken word. Moreover, given the typically short duration of phones, we wanted to minimize the possibility of introducing artificial speech modifications (or distortions), which might distract listeners and their responses to the stimuli.

Following this design decision, it was important to select pitch and speech tempo modification limits to within two standard deviations of typical speech. For speech tempo, two ratios of modified to normal speech tempo were developed, where each fell within two standard deviations below ($\frac{1}{2}$) and above ($\frac{3}{2}$) a mean maximum coefficient of variation of $\approx 25\%$ [23]. In the pitch domain, a mean of 25% variation equates to 3 semitones or 300 cents. Therefore, to be consistent with speech tempo, we opted to raise and lower pitch by 600 cents, as two standard deviations - in this case, 50% - represents $\frac{1}{2}$-octave. Thus, in addition to studying whether pitch and speech tempo modifications influenced perceptual SID performance, we also examined any effects of modification direction (increasing or decreasing).

To modify speech recordings, the mPraat Matlab toolbox [26] was used to load speech recordings and speech-analysis Praat [27] Textgrid files into MATLAB. Next the number of words in each speech recording were calculated and distinguished the first 75% of the words (rounded to the nearest integer) from the last 25%, which remained un-modified. The mean duration of words was 0.28 $\pm$ 0.15 s ranging from 0.03 to 0.89 s. Each word in the first 75% was given a percentage of modification, such that the first word was given 100% and the last 0% (linear ramp). The corresponding audio was extracted and modified using the "Waveform Similarity Overlap-add" function *wsolaTSM*, which is available in the MATLAB TSM toolbox [28]. Each audio extract was modified based on the type of modification (pitch or speech tempo) and percentage of modification. For example, if the word "bise" ("kiss") had been shifted up in pitch with a percentage of 45%, then it would be shifted up 270 cents. All audio extracts were then concatenated together. Following modification and concatenation processing, the mean duration of words was 0.28 $\pm$ 0.15 s ranging from 0.03 to 1.04 s. The mean duration of each modified speech recording was 10.3 $\pm$ 1.3 s.

2.4. Participants

39 people (31 female and 8 male) participated in the perceptual study (mean age 28.1 $\pm$ 13.6 years). All participants were native-French speakers and reported good hearing. All participants consented to voluntary participation in the study and were informed of their right to withdraw at any time. They were compensated for their time.

2.5. Experimental setup

2.5.1. Technical setup

Participants completed trials on desktop computers at CEP-LPL. Throughout the study participants wore Superlux HD 681B headphones. Prior to testing, participants listened to a speech recording and adjusted the volume to their comfort.

2.5.2. Familiarization trials

In order to task naive listeners with identifying speakers whose speech characteristics were modified, it was necessary to provide them first with a period of familiarization. Rather than reusing in-set speaker stimuli to familiarize participants (see Section 2.2) and thus potentially introducing memory bias, a separate set of speech recordings were used. This stimuli contained spontaneous speech produced by the same in-set speakers in a separate recording session. In this recording session, the speakers were asked to describe their educational and professional experiences. Although the speech form differed between familiarization and testing stimuli, evidence has shown that spontaneous speech is more natural and rich in prosody [29], which listeners might take advantage of when characterizing speakers. Eight speech recordings were selected per in-set speaker and were edited to remove any formal identifiers. The duration of the training speech recordings ranged from 3.9 to 6.9 s (mean duration: 5.1 $\pm$ 0.8 s).

First participants familiarized themselves with the speech characteristics of each in-set speaker (4) by clicking on a speech recording located below a photo portrait of the speaker on a Microsoft Powerpoint file (limitless).

Next participants were tasked to complete a series of trials programmed in Perceval [30]. These trials contained unmodified (normal) spontaneous speech recordings of the in-set speakers. At the start of each trial, a spontaneous speech recording belonging to an in-set speaker was played. In addition photo portraits of each speaker were aligned on the screen. Participants were tasked to select the photo portrait that corresponded with the speech recording. After each response, incorrect photo portraits were removed from the screen, while the correct one remained for 5 s before starting the next trial. Following a series of eight trials, participants received a score (%) that reflected their performance accuracy. They were required to receive a score of 100% for two rounds,

2.5.3. Modified speech trial design

Participants were tasked to complete two rounds of 48 "go/no-go" trials programmed in Lancelot [30] with modified speech recordings. For each "go/no-go" trial, participants were presented a speech recording and a photo portrait. If they believed there was a correspondence between the photo portrait and the speech recording, then they considered the trial a "go" and responded by clicking on a button placed in front of them. If they believed there was a discontinuity between the two stimuli, then it was considered a "no-go" trial and did not register a response. To indicate the start of a trial, participants were presented a short "beep" generated by a sinusoidal oscillator (frequency: 500 Hz; duration: 0.8 s).

For each round, each in-set and out-of-set speaker (8 total) was presented six times, three per modification direction (1:1 ratio of "go" to "no-go" trials). Each round corresponded to a different modification type (pitch or speech tempo), where the order alternated between participants (balanced). Between each round, participants re-examined the Powerpoint file so as to refamiliarize themselves with the normal speech characteristics of the in-set speakers.

2.6. Data processing

To evaluate perceptual SID performance, accuracy was calculated, such that if participants correctly responded during "go" trials (*true positive*) by correctly identifying in-set speaker with their corresponding face or did not respond during "no-go" trials (*true negative*) i.e. found the in-set speaker face did not correspond with the out-set speaker voice, then they received a score of 1.0. They received a score of 0.0 for all *false positive* and *false negative* responses.

To evaluate the effects of speech modifications on the time required to respond per trial, a percentage of modification metric was developed. First, "go" trials where participants correctly responded were identified. Next, given the time of response, the corresponding location in the speech recording was identified. Finally a percentage of modification was calculated based on the amount of modified speech at this location. For example, a participant who correctly responded when pitch was elevated 300 cents would have a percentage of modification of 50%. In addition to the fact that all speech files had different durations, this metric allowed us to neutralise the effects of speech tempo modifications, where the direction either shortened or elongated speech and subsequently the recording's duration.

For all outcome variables, Repeated Measures ANOVAs were carried out with Greenhouse-Geisser adjustments. We reported main effects on speakers and speech modification types, as well as their interactions. Because the speech recordings began with a maximum percentage of modification (100%) that was either above or below normal speech, direction (shifted up or down) as dependent on speech modification type. Thus, interactions between modification and direction types were reported. Where main effects and interactions were detected, post-hoc Bonferroni-adjusted t-tests were carried out (α = 0.05).

2.7. Preliminary results

Participants had a mean accuracy of 73.5 $\pm$ 6.5% ranging from 58.3% to 83.3%. Their mean response time for correct "go" responses was 5.78 $\pm$ 1.45 s with a range from 2.98 s to 8.44 s. Normal distribution functions were fitted with accuracy and response time metrics, which showed all participant performances fell within two standard deviations of the mean. Pearson correlation procedures revealed no significant relationship between accuracy and response times (ρ = 0.08, $p > 0.05$).

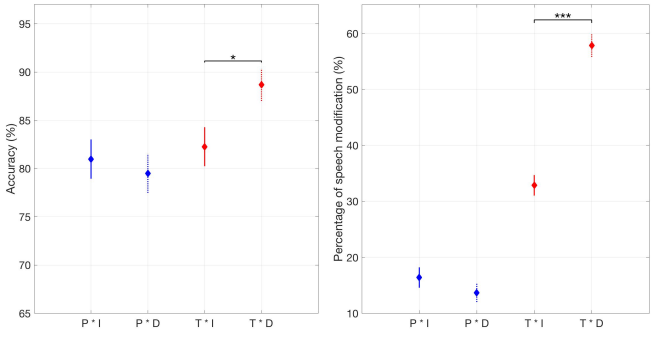
3. Results

3.1. Accuracy

Main effects on accuracy were observed for speakers $F_{3,114}$ = 6.42, $p < 0.01$, η_p^2 = 0.14 and modification types $F_{1,38}$ = 6.66, $p < 0.05$, η_p^2 = 0.14, as well as interactions between speakers * modification types $F_{3,114}$ = 3.0, $p < 0.05$, η_p^2 = 0.07 and modification * direction types $F_{1,38}$ = 4.67, $p < 0.05$, η_p^2 = 0.11. Post-hoc tests on main effects revealed participants found speaker LG024 (74.6 $\pm$ 2.4%) significantly more difficult to identify in comparison to speakers LG005 (85.0 $\pm$ 1.8%, $p <$ 0.05) and LG013 (87.0 $\pm$ 1.7, $p < 0.05$), but not LG016 (84.8 $\pm$ 1.7, $p > 0.05$). They also had significant difficulty when presented speech recordings with modified pitch (P) (80.2 $\pm$ 1.4%) in comparison to modified speech tempo (T) (85.5 $\pm$ 1.3%), $p < 0.05$. Post-hoc tests on speaker * modification interactions revealed participants found significant differences between modification types for speakers LG013 (P: 82.1 $\pm$ 2.8%; T: 91.9 $\pm$ 1.7%), LG016 (P: 80.3 $\pm$ 2.6%; T: 89.3 $\pm$ 2.2%), and LG024 (P: 72.7 $\pm$ 3.5%; T: 76.5 $\pm$ 3.3%), $p < 0.05$, however, not LG005 (P: 85.9 $\pm$ 2.3%; T: 84.2 $\pm$ 2.8%). Post-hoc tests on modification * direction interactions revealed participants improved performance when speech recordings had decreased speech tempo modifications (88.7 $\pm$ 1.7%) in compar-

ison to increased speech tempo modifications (82.3 $\pm$ 2.0%). No significant effects of direction were observed for pitch modifications (P*D: 79.5 $\pm$ 2.0%; P*I: 81.0 $\pm$ 2.0%), $p > 0.5$ (Fig 1 - Left).

Figure 1: *Interactions between modification and direction types: Accuracy (Left) and Percentage of modification (Right*



3.2. Percentage of modification

A main effect on the percentage of modification was observed on modification type $F_{1,31}$ = 121.23, $p < 0.001$, η_p^2 = 0.8, but not on speakers, $p > 0.05$. In addition, an interaction between modification * direction types $F_{1,31}$ = 77.84, $p < 0.001$, η_p^2 = 0.72 was observed. Despite increased percentages of modification, participants were able to respond accurately when presented speech recordings with speech tempo modifications (45.4 $\pm$ 1.6%) in comparison to those with pitch modifications (15.0 $\pm$ 1.2%), $p < 0.001$. Post-hoc tests on modification * direction interactions revealed participants were able to respond accurately when presented speech recordings with decreased speech tempo modifications (57.9 $\pm$ 2.1%) in comparison to increased speech tempo modifications (32.9 $\pm$ 1.8%) (Figure 1 - Right). No significant effects of direction were observed for pitch modifications (P*D: 16.4 $\pm$ 1.8%; P*I: 13.7 $\pm$ 1.7%), $p >$ 0.5.

4. Discussion

The goal was to examine whether modified speech affected perceptual SID performance by naive listeners. The findings showed participant performance was significantly affected by the presence of speech recordings containing pitch modifications in comparison to those with speech tempo modifications. Although this observation was anticipated, an important finding was that listeners had significantly lower tolerances for pitch modifications, as evidenced by the percentage of modification for interactions between pitch * decrease (16.4 $\pm$ 1.8%) and pitch * increase (13.7 $\pm$ 1.7%). These percentages equated to -98.4 and +82.2 cents, respectively, which are both $<$ 1 semitone. Unlike speech tempo modifications, no direction influenced the effects of pitch modifications, which suggests that elevated or reduced pitch modifications on speech have similar effects of perceptual SID performance.

A significant effect of direction was observed on speech tempo modifications, where participants enhanced performance when presented speech recordings containing decreased speech tempo modifications (88.7 $\pm$ 1.7%). A similar trend was observed in the percentage of modification results, which revealed participants were able to respond not only correctly but faster when presented decreased speech tempo modifications (57.9 $\pm$ 2.1%) despite the increased percentage of modifications. These initial findings suggest that by decreasing speech tempo, listeners had more time to focus and associate speech sounds with specific speakers. Unlike decreased speech tempo modifications, increased speech tempo modifications had effects that were similar to pitch modifications. These results suggest that by increasing speech tempo, each word's duration was compressed, which, in turn, saturated information processing. Linguistic content is less understandable with increased speech tempo modifications [31]. Therefore listeners were destabilized and may have found it more difficult to associate voices with specific speakers. However, this destabilization was much less in comparison to experiences during pitch modifications, as listeners were able to tolerate increased speech tempo modifications as evidenced by 32.9 $\pm$ 1.8%, which was slightly over the $\approx$ 25% mean maximum coefficient of variation reported in [23].

The effects of modified speech on speakers showed participants were less accurate when presented speaker LG024 in comparison to LG005 and LG013. Table 1 reveals LG024 had a much smaller mean ED with the out-of-set speakers (4.2 $\pm$ 2.1) in comparison to LG005 (12.3 $\pm$ 2.8) and LG013 (11.8 $\pm$ 2.9). Thus from the outset, participants were predisposed to have more difficulty distinguishing LG024 from out-of-set speakers, which may have been further compromised by the task of making distinctions between modified speech recordings. Another important take-away was that speech modifications had no significant effect on any one speaker, as evidenced by our percentage of modification results. This finding suggests modifications affected speakers equally and did not bias certain speech characteristics, given their F0 and speech tempo ranges of 101 to 138 Hz and 12.3 to 15.3 pho/s, respectively.

While findings on the significant effects of modified pitch and direction on speech tempo modifications were reported, it was important to highlight inter-listener variability, where accuracy and response times ranged from 58.3% to 83.3% and 2.98 s to 8.44 s. Notably the lower limit verges on random responses (50%), which suggests some participants found the task more difficult than others. Despite completing the same trials with similar stimuli and having the same-level of familiarity with speakers, naive listeners exhibited a sizeable range of variability.

5. Conclusions

This study examined the effects of modified speech on perceptual SID performance by naive listeners. The following were major take-away messages from the study: (1) participants were more affected by modified pitch in comparison to modified speech tempo, which suggests the latter was used less to distinguish speakers; (2) the direction of the modification played a significant role on speech tempo modifications, but not for pitch modifications; and (3) there was strong between-listener variability, which supports evidence that listeners model speakers differently.

6. Acknowledgements

This work was funded by the French National Research Agency (ANR) under the VoxCrim project (ANR-17-CE39-0016). We thank Estelle Chardenon for her help developing stimuli.

7. References

[1] S. V. Stevenage, "Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings," *Neuropsychologia*, vol. 116, pp. 162–178, 2018, special Issue: Familiar Voice Recognition. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002839321730252X

[2] K. v. K. Samuel R. Mathias, "How do we recognise who is speaking?" *Frontiers in Bioscience-Scholar*, vol. 6, no. 1, pp. 92–109, 2014.

[3] J. Kreiman and D. Van Lancker Sidtis, *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*, 04 2011.

[4] S. Schweinberger, H. Kawahara, A. Simpson, V. Skuk, and R. Zäske, "Speaker perception," *Wiley interdisciplinary reviews. Cognitive science*, vol. 5, 01 2014.

[5] S. Mattys, M. Davis, A. Bradlow, and S. Scott, "Speech recognition in adverse conditions: A review," *Language and Cognitive Processes - LANG COGNITIVE PROCESS*, vol. 27, pp. 953–978, 09 2012.

[6] N. Lavan, L. Burston, and L. Garrido, "How many voices did you hear? natural variability disrupts identity perception from unfamiliar voices," *British Journal of Psychology*, vol. 110, 09 2018.

[7] H. Smith, T. Baguley, J. Robson, A. Dunn, and P. Stacey, "Forensic voice discrimination: The effect of speech type and background noise on performance," *Applied Cognitive Psychology*, vol. 33, 10 2018.

[8] I. Johnsrude, A. Mackey, H. Hakyemez, E. Alexander, H. Trang, and R. Carlyon, "Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice," *Psychological science*, vol. 24, 08 2013.

[9] D. Van Lancker and J. Kreiman, "Voice discrimination and recognition are separate abilities," *Neuropsychologia*, vol. 25, no. 5, pp. 829–834, 1987.

[10] P. Bestelmeyer and C. Mühl, "Individual differences in voice adaptability are specifically linked to voice perception skill," *Cognition*, vol. 210, p. 104582, 05 2021.

[11] O. Baumann and P. Belin, "Perceptual scaling of voice identity: Common dimensions for different vowels and speakers," *Psychological research*, vol. 74, pp. 110–20, 12 2008.

[12] 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.

[13] C. LaRiviere, "Some acoustic and perceptual correlates of speaker identification," in *Proceedings of the Seventh International Congress of Phonetic Sciences*, 1971, pp. 558–564.

[14] P. Bricker and S. Pruzanski, "Effects of stimulus content and duration on talker identification," *J Acoust Soc Am*, vol. 40, pp. 1442–1449, 01 1966.

[15] I. Pollack and J. Pickett, "On the identification of speakers by voice," *Journal of The Acoustical Society of America*, vol. 26, 05 1954.

[16] R. Roebuck and J. Wilding, "Effects of vowel variety and sample length on identification of a speaker in a line-up," *Applied Cognitive Psychology*, vol. 7, pp. 475–481, 11 1993.

[17] H. Traunmüller and A. Eriksson, "The frequency range of the voice fundamental in the speech of male and female adults," Department of Linguistics, University of Stockholm, Tech. Rep., 1994.

[18] J. Harrington, S. Palethorpe, and C. Watson, "Age-related changes in fundamental frequency and formants: A longitudinal study of four speakers," in *Proceedings of INTERSPEECH 2007*, 08 2007, pp. 2753–2756.

[19] R. Bezooijen, "Sociocultural aspects of pitch differences between japanese and dutch women," *Language and speech*, vol. 38 (Pt 3), pp. 253–65, 07 1995.

[20] H. Quené, "On the just noticeable difference for tempo in speech," *Journal of Phonetics*, vol. 35, pp. 353–362, 01 2001.

[21] W. Van Dommelen, "The contribution of speech rhythm and pitch to speaker recognition," *Language and Speech*, vol. 30, pp. 325–338, 10 1987.

[22] M. Ajili, J.-F. Bonastre, W. Kheder, S. Rossato, and K. J, "Homogeneity measure impact on target and non-target trials in forensic voice comparison," in *INTERSPEECH*, 2017, pp. 2844–2848.

[23] E. Chardenon, C. Meunier, and C. Fougeron, "Variations de débit articulatoire sur un corpus conversationnel avec différents types d'interactions," in *Proceedings of the 34th Journées d'Etudes sur la Parole*, 2022.

[24] A. Chanclu, L. Georgeton, C. Fredouille, and J.-F. Bonastre, "Ptsvox: une base de données pour la comparaison de voix dans le cadre judiciaire," in *6e conférence conjointe Journées d'Études sur la Parole*, 2020, pp. 73–81.

[25] A. Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, pp. 1917–30, 05 2002.

[26] T. Boril and R. Skarnitzl, "Tools rpraat and mpraat," vol. 9924, 09 2016, pp. 367–374.

[27] P. Boersma, "Praat, a system for doing phonetics by computer," vol. 5, no. 9/10, pp. 341–345, 2001.

[28] J. Driedger, M. Müller, and S. Ewert, "Improving time-scale modification of music signals using harmonic-percussive separation," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 105–109, 2014.

[29] E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care," in *9th European Conference on Speech Communication and Technology*, 09 2005, pp. 1781–1784.

[30] C. Andre, A. Ghio, C. Cavé, and B. Teston, "Perceval: a computer-driven system for experimentation on auditory and visual perception," in *International Congress of Phonetic Sciences*, 06 2007, pp. 1421–1424.

[31] K. Perrachione, Tyler, "Speaker recognition across languages," 2017. [Online]. Available: https://open.bu.edu/handle/2144/23877