IDEAR: A SPEECH DATABASE OF IDENTITY-MARKED, CLEAR, AND READ SPEECH

Valeriia Perepelytsia[1], Leah Bradshaw[1], and Volker Dellwo[1]

[1] Department of Computational Linguistics, University of Zurich, Zurich, Switzerland
valeriia.perepelytsia@uzh.ch, leah.bradshaw@uzh.ch, volker.dellwo@uzh.ch

ABSTRACT

Speakers adapt their speech in response to communicative context and listeners' needs. For example, when talking to hard of hearing or non-native listeners, or in the presence of background noise speakers make their speech more intelligible by speaking clearly. However, it is not clear which adaptations speakers will make when the goal is not intelligibility, but voice recognition. In this paper, we describe a speech database collected using a novel Wizard of Oz technique to investigate speakers' vocal adaptations when they are prompted to sound either more intelligible (clear speech) or more recognizable (identity marked speech). We recorded 39 speakers interacting with a mock speech recognizer, which repeatedly misunderstands speech, and a mock speaker recognizer, which misrecognizes the speakers' voice. We also collected read speech which served as a baseline. All recordings have been orthographically transcribed, forced aligned and manually corrected.

Keywords: speaker recognition, clear speech, speaker characteristics.

1. INTRODUCTION

Speakers dynamically adapt their speech depending on the communicative intent and/or perceptual difficulties on the listeners' side. For example, when talking to hearing-impaired individuals [1], children with and without learning disabilities [2], non-native listeners [3], when communicating under varying levels of background noise [4], or when talking to computers or voice-AI assistants [5] speakers adopt a mode of talk which can be broadly defined as *clear speech*, an intelligibility-enhancing speaking style. Generally, this clear speech is characterized by decreased speaking rate, longer segments, more frequent and longer pauses, increased intensity, and expanded vowel space compared to conversational speech [1], [6].

In addition to varying the degree of speech intelligibility, there is evidence to suggest that speakers can control the degree to which their voice is recognizable, for example, by means of voice disguise [7]–[10]. When disguising their voices, speakers tend to manipulate either the inborn speaker-specific information such as voice fundamental frequency [8] and/or acquired information such as regional accent. This suggests that speakers have some intuition about the speaker-specific information important for their and others' voice identity and are aware of some strategies for changing it.

Although it is evident that speakers can conceal their identity and make themselves less recognizable, it is not clear whether they can move in an opposite direction, namely, enhancing the recognizability of their voices. The ability to manipulate voice individuality and thus enhancing or reducing its recognizability is an important part of vocal communication in humans and many other animals. In human communication, the information about who is speaking is crucial for structuring and understanding speech content [11]. In animals, discriminating among individuals is vital for offspring and mate recognition, cooperative behaviors, and social hierarchies [12].

In this paper, we describe a database collected to investigate acoustic differences in speaking styles when speakers are presented with two types of communicative pressures: (1) being more intelligible (clear speech); and (2) being more recognizable (identity marked speech). Because clear speech is aiming at intelligibility, speakers tend to enhance the canonical properties of segments and suprasegmental features. Thus, speaker-specific properties should be mostly suppressed (see discussion in [11]). Revealing voice identity, on the other hand, implies that speaker-specific anatomical and acquired features should be enhanced. With our database, we want to investigate if speakers reveal more speaker-specific properties in identity-marked speech and whether this leads to their voices being better recognizable under such circumstances. Below we describe the speaker population, the experimental paradigm, the speech tasks, as well as the post-processing of audio files. The motivation for collecting this database is twofold: first, exploring to what extent speakers can enhance the speaker-specific information in their speech, and thus become more recognizable. Also, previous studies considering human-computer interactions (HCI) focus on speech recognition systems [13], [14], while voice-AI assistants are increasingly being used for alternative tasks, such as voice identification (for example, in banking). To our knowledge, this experiment is a first attempt to explore voice-AI-directed speech on a wider scale.

2. DESIGN AND METHOD

2.1. General setup

To collect speech data in this experiment, we used the Wizard of Oz (WOz) paradigm, which has been widely used in HCI research [15]–[19]. In WOz experiments, participants interact with a computer system that they believe to be autonomous, but which in fact is being operated by an unseen experimenter (i.e., the wizard). WOz paradigm means participants receive seemingly relevant feedback from the system being tested but full experimental control can be maintained. It also allows eliciting different speech modes without explicit instructions for speakers to modify their speech in a certain way.

In our experiment, participants were asked to interact with two mock automatic systems, which we were in the process of developing and testing. The speakers completed three tasks. The first task was reading aloud 34 sentences. Participants were informed that the purpose of this task was to obtain a sample of their speech on which our systems would train. Speakers would then interact with our *speaker recognition* system, Verifico, modelled after the kinds of voice identification technology used in banking. Verifico was aiming to recognize/identify them from their voice sample. After this, speakers would interact with our *speech recognition* system, Vicky, which was similar to Amazon's Alexa or Apple's Siri. Vicky was attempting to correctly understand the speech of the participant.

To the participants knowledge, both Verifico and Vicky were genuine pieces of software which were developed due to a lack of existing systems for Swiss German speakers. However, for the purpose of greater experimental control, the feedback from both systems was fully designed and regulated by the experimenters.

2.2. Speakers

We recorded 39 speakers (mean age = 25.9 years, range = 19–34, SD = 3.36; 20 female) from the University of Zurich student population. All were native speakers of Swiss German. 67 % of speakers spoke Zurich dialect of Swiss German as their native dialect, while the rest of the speakers spoke other dialects of Swiss German. None of the speakers reported any history of speech, language, or hearing disorders. All gave their written consent and received a monetary compensation.

2.3. Recording conditions and procedure

Recordings were made in a sound-attenuated booth at the *Linguistic Research Infrastructure* (LiRI) laboratory at the University of Zurich using a Røde NT1 microphone. Recordings were made directly to disk in WAV format at 44.1 kHz sampling rate, 16 kbit/sec bitrate using Pro Tools software [20].

Each speaker participated in one recording session which lasted approximately 60 minutes. Prior to the recording, speakers were briefed about the procedure and tasks of the experiment. The experimenters informed the speakers that they will be communicating with two automatic systems: a speech recognizer and a speaker recognizer. The experimenters explained the basic principles of how the two systems work, as well as the key differences between the systems. Namely, that the speech recognizer is aiming for *speech recognition*, while the speaker recognizer is targeting *voice identification*.

Speakers were told that both systems are still in the development phase and that they have been already pretrained on a large number of other speakers, but that they still require calibration to improve their performance. They were also told that due to large variability between speakers the systems might make errors: the speaker recognizer might confuse the current speaker with the other speakers from the database on which it was pretrained, and the speech recognizer might misrecognize different words in a produced utterance.

After the briefing, the recording began. During the recording, the speaker sat in front of a computer screen on which prompts were presented via the Gorilla experiment builder [21]. During the speech tasks, the speakers received feedback from the mock automatic systems both visually on the screen and aurally via the loudspeakers inside the booth.

2.4. Speech tasks

The recording session consisted of three speech tasks, which the speakers were instructed to do in Swiss Standard German [22]: the read sentences task, followed by the speaker recognizer task for eliciting identity marked speech, and lastly the speech recognizer task for eliciting clear speech. We opted for Swiss Standard German since common voice-AI assistants, such as Siri or Alexa, do not yet exist for Swiss German, and therefore this was a plausible study motivation for our participants.

The order of tasks was fixed: all speakers started with the read sentences task, followed by the speaker recognizer task, and, finally, the speech recognizer task. We chose this fixed task order as speakers were far more familiar with clear speech strategy compared to identity-marked speech strategy, because of more experience with speech recognition tools such as Apple's Siri or Amazon's Alexa. Therefore, we decided to present speakers with the speaker recognizer task before the speech recognizer task to avoid transfer effects from one task to another.

2.3.1. Read sentences task

In the first task, the speakers read 34 phonetically rich sentences presented in a random order. All sentences had SVO structure and were 5 words in length. The speakers were informed that both systems were already pretrained on the existing databases and their read speech would be used to enrol them into the datasets. This task was included as a baseline for future acoustic analyses.

2.3.2. Speaker recognizer task

In the second task, participants were interacting with the mock *speaker* recognizer. The task started with the brief interaction between the participant and the system. The system introduced itself as Verifico and asked the participant to select their gender and indicate their name. It then assigned an avatar to the speaker. This avatar was shown whenever the system "correctly recognized" the participant during the task. This introduction was included to give the participant the impression that the system was attempting to learn something about them and their identity.

We used text-to-speech (TTS) software to generate Verifico's utterances (available at https://ttsmp3.com). This software is part of Amazon Polly, a TTS cloud service supporting multiple languages and a variety of lifelike voices. We chose a male voice for our speaker recognizer system.

After the introduction, speakers did several practice trials to make sure they understood the goal of the task, which was to make the system recognize their *voice* correctly. The practice trials were excluded from subsequent analyses and are not available in the database. After the practice rounds, the real task began. Stimuli for this task consisted of the same 34 sentences that the participants had previously produced in the read sentences task.

In each trial (n=34), speakers would say the sentence prompted on the screen and wait for the system's feedback. Each trial had three feedback options. The first option was a correct recognition, whereby the speaker's avatar and their name appeared on the screen. The second option was a single misrecognition, whereby the avatar of another speaker and a different person's name appeared on the screen; in this case, the speaker would have to read the sentence again and would receive a correct recognition after that. The third option was a double misrecognition, whereby the speaker would read a sentence and receive two misrecognitions before being correctly recognized. All three feedback options were equally probable. In each trial, one feedback option was randomly selected by the Gorilla randomizing algorithm. The schematic overview of each trial is given in Figure 1.

During this task, speakers produced on average 67 utterances for all 34 trials in total (SD = 4.07 utterances, range = 60–75). On average, in 34.7% of the trials, speakers were correctly recognized after the first sentence production (SD = 7.5%, range = 17.6–50%), in 32.4% of the trials they received a single misrecognition (SD = 8.8%, range = 17.6–52.9%), and in the remaining number of trials – a double misrecognition (SD = 7.5%, range = 17.6–47%).

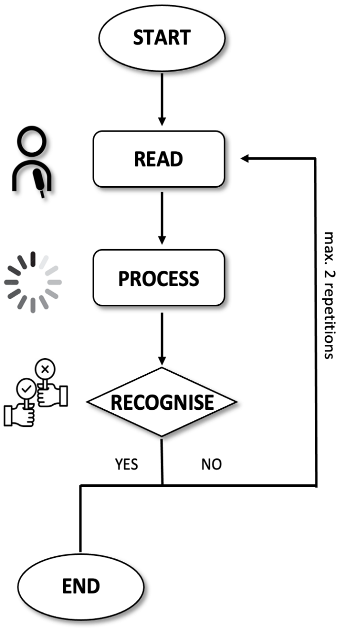

Figure 1: Flow chart showing trial structure in speaker and speech recognizer tasks. The speakers would read the prompted sentence and receive the system's feedback, which consisted of three equally probable options: (1) correct recognition; (2) a single misrecognition; and (3) a double misrecognition.

2.3.3. Speech recognizer task

In the third task, speakers were interacting with the mock *speech* recognizer named Vicky. We chose female voice for the speech recognizer to highlight the difference between the two systems in this experiment. We generated Vicky's utterances using the same TTS software as in the Verifico task described above.

This task also started with Vicky introducing herself to the participants. The introduction was followed by a similar round of practice trials as in the previous task. The practice trials were included to ensure the speakers understood the difference between the two systems and how to complete this new task, which now consisted of making the system recognize their utterances correctly. In addition, the introduction to the new system and the practice trials served as a buffer to clearly separate the two systems and two tasks for the participants. The practice trials were again excluded from subsequent analyses. After this, the real task began.

As with the speaker recognizer task, the same 34 sentences were used as prompts for what the speakers should say to the system. Each trial (n=34) was structured similarly to the trials in the speaker recognizer task (see Figure 1), however, the feedback given to the speakers by the system differed. Speakers would read a sentence prompted on the screen and wait for Vicky's feedback. The feedback consisted of three equally probable options, which were randomly selected in each trial. The first option was a correct speech recognition, whereby the whole utterance was recognized correctly. The second option was a single misrecognition, whereby two words out of five in a produced utterance were "misheard" by the system. In this instance, speakers would be requested to produce the same sentence again and would be correctly understood by Vicky following this repetition. The final option was a double misrecognition, whereby the system would offer two misrecognitions before a correct recognition. In this instance, each misrecognition would contain two of five words in an utterance "misheard" by the system, and these "misheard" words would be different in each of the two misrecognitions. In the latter case, the speakers would repeat the sentence three times in total before a correct recognition. For instance, a possible double misrecognition scenario may be:

Original sentence:
German: *Die Richterin verliest das Urteil.*
English: 'The judge reads out the verdict.'
Misrecognition 1:
German: *Die Mieterin verliest den Hauptteil.*
English: 'The tenant reads out the main part.'
Misrecognition 2:
German: *Die Freundin vergisst das Urteil.*
English: 'The girlfriend forgets the verdict.'

The misrecognized words were selected based on the acoustic similarity with the target words to ensure that Vicky's misrecognitions were plausible and acceptable to the speakers. Misrecognition of single words might have led to local intelligibility adjustments (as shown by [13]), while we wanted speakers to produce intelligibility adjustments on the utterance level, and not on the word level. Therefore, we opted for misrecognizing two words out of five in each misrecognized utterance. This technique was also adopted to persuade speakers that word-level intelligibility adjustments would not assist the system with recognizing their speech.

During this task, speakers produced on average 67.5 utterances for all 34 trials in total (SD = 5.08 utterances, range = 51–80). In 33.3% of the trials, speakers were correctly recognized after the first sentence production (SD = 7.5%, range = 20.6–52.9%), in 33% of the trials they received a single misrecognition (SD = 7.4%, range = 17.6–52.9%), and in the remaining number of trials – a double misrecognition (SD = 7.5%, range = 17.6–55.9%).

For the preliminary results of acoustic analyses of the collected speech data, see [23].

2.5. Transcription and alignment

2.5.1. Orthographic transcription

All recordings were manually segmented and labelled, before individual sentence productions were extracted using a custom Praat [24] script. Sentence-level orthographic transcriptions were created for all utterances in the form of Praat TextGrids, whereby the first TextGrid tier contained the corresponding sentence, e.g., "Die Pflanze verliert ihre Blätter." (English: "The plant loses its leaves.")

2.5.2. Automatic forced alignment

All utterances were subjected to automatic forced alignment with the help of Montreal Forced Aligner (MFA) [25]. We chose MFA because it has previously shown better performance compared to other widely used forced aligners such as FAVE and MAUS [26]. For each utterance, MFA uses the wav file and the corresponding orthographic transcription in TextGrid format to generate word and phone level alignments. Although the current database contains speech in Swiss Standard German, the alignments were created using MFA Standard German dictionary and acoustic model, which were slightly adapted for the content of the dataset. MFA has been shown to perform well even when the dialect in the dataset to be aligned differs from the one on which the acoustic model was created [26].

All forced alignments were manually corrected by two trained phoneticians (the first and second authors): word boundaries were corrected, and vowel onsets and offsets were defined based in the presence of periodicity and higher formant structure.

6. ACKNOWLEDGEMENTS

This work has been supported by the Swiss National Science Foundation grant #185399. The recordings were carried out at the Linguistic Research Infrastructure laboratory of the University of Zurich.

8. REFERENCES

[1] M. A. Picheny, N. I. Durlach, and L. D. Braida, ‘Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech’, *J. Speech Lang. Hear. Res.*, vol. 29, no. 4, pp. 434–446, Dec. 1986, doi: 10.1044/jshr.2904.434.

[2] A. R. Bradlow, N. Kraus, and E. Hayes, ‘Speaking clearly for children with learning disabilities: Sentence perception in noise’, *J. Speech Lang. Hear. Res.*, vol. 46, no. 1, pp. 80–97, Feb. 2003, doi: 10.1044/1092-4388(2003/007).

[3] M. Uther, M. A. Knoll, and D. Burnham, ‘Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech’, *Speech Commun.*, vol. 49, no. 1, pp. 2–7, Jan. 2007, doi: 10.1016/j.specom.2006.10.003.

[4] K. L. Payton, R. M. Uchanski, and L. D. Braida, ‘Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing’, *J. Acoust. Soc. Am.*, vol. 95, no. 3, pp. 1581–1592, Mar. 1994, doi: 10.1121/1.408545.

[5] D. Burnham, S. Joeffry, and L. Rice, ‘Computer- and human-directed speech before and after correction’, presented at the Proceedings Of The 13Th Australasian International Conference On Speech Science And Technology, Melbourne, Australia, 2010, pp. 13–17.

[6] R. Smiljanić and A. R. Bradlow, ‘Production and perception of clear speech in Croatian and English’, *J. Acoust. Soc. Am.*, vol. 118, no. 3, pp. 1677–1688, Sep. 2005, doi: 10.1121/1.2000788.

[7] T. Kitamura, ‘Acoustic analysis of imitated voice produced by a professional impersonator’, presented at the INTERSPEECH 2008, 2008.

[8] I. Hove and V. Dellwo, ‘The effect of voice disguise on f0 and on the formants’, in *Proceedings of IAFPA*, 2014.

[9] A. Eriksson, ‘The disguised voice: Imitating accents or speech styles and impersonating individuals’, in *Language and Identities*, C. Llamas and D. Watt, Eds., Edinburgh University Press, 2010, pp. 86–96.

[10] A. Eriksson and P. Wretling, ‘How flexible is the human voice? A case study of mimicry’, presented at the Fifth European Conference on Speech Communication and Technology, 1997.

[11] V. Dellwo, E. Pellegrino, L. He, and T. Kathiresan, ‘The dynamics of indexical information in speech: Can recognizability be controlled by the speaker?’, *AUC Philol.*, vol. 2019, no. 2, pp. 57–75, Oct. 2019, doi: 10.14712/24646830.2019.18.

[12] E. A. Tibbetts and J. Dale, ‘Individual recognition: It is good to be different’, *Trends Ecol. Evol.*, vol. 22, no. 10, pp. 529–537, Oct. 2007, doi: 10.1016/j.tree.2007.09.001.

[13] M. Cohn and G. Zellou, ‘Prosodic differences in human- and Alexa-directed speech, but similar local intelligibility adjustments’, *Front. Commun.*, vol. 6, p. 675704, Jul. 2021, doi: 10.3389/fcomm.2021.675704.

[14] K. Maniwa, A. Jongman, and T. Wade, ‘Acoustic characteristics of clearly spoken English fricatives’, *J. Acoust. Soc. Am.*, vol. 125, no. 6, pp. 3962–3973, Jun. 2009, doi: 10.1121/1.2990715.

[15] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, ‘Wizard of Oz studies — why and how’, *Knowl.-Based Syst.*, vol. 6, no. 4, pp. 258–266, Dec. 1993, doi: 10.1016/0950-7051(93)90017-N.

[16] D. DeVault, J. Mell, and J. Gratch, ‘Toward natural turn-taking in a virtual human negotiation agent’, presented at the AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction., 2015.

[17] L. Gauder, M. Reartes, R. H. Gálvez, Š. Beňuš, and A. Gravano, ‘Testing the effects of acoustic/prosodic entrainment on user behavior at the dialog-act level’, in *Speech Prosody 2018*, ISCA, Jun. 2018, pp. 374–378. doi: 10.21437/SpeechProsody.2018-76.

[18] S. Oviatt, C. Darves, and R. Coulston, ‘Toward adaptive conversational interfaces: modeling speech convergence with animated personas’, *ACM Trans. Comput.-Hum. Interact.*, vol. 11, no. 3, pp. 300–328, Sep. 2004, doi: 10.1145/1017494.1017498.

[19] I. Gessinger, B. Möbius, N. Fakhar, E. Raveh, and I. Steiner, ‘A Wizard-of-Oz experiment to study phonetic accommodation in human-computer interaction’, presented at the International Congress of Phonetic Sciences (ICPhS), Melbourne, 2019, pp. 1475–1479. [Online]. Available: https://www.assta.org/proceedings/ICPhS2019/papers/ICPhS_1524.pdf

[20] E. Brooks and P. Gotcher, ‘Pro Tools’. Avid Audio. Accessed: Nov. 10, 2022. [Online]. Available: https://www.avid.com/pro-tools

[21] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, ‘Gorilla in our midst: An online behavioral experiment builder’, *Behav. Res. Methods*, vol. 52, no. 1, pp. 388–407, Feb. 2020, doi: 10.3758/s13428-019-01237-x.

[22] I. Hove, *Die Aussprache der Standardsprache in der deutschen Schweiz*. DE GRUYTER, 2002. doi: 10.1515/9783110919936.

[23] L. Bradshaw, V. Perepelytsia, and V. Dellwo, ‘Vocal effort in human interactions with voice-AI’, in *Proceedings of ICPhS*, 2023 [accepted].

[24] P. Boersma and D. Weenink, ‘Praat: doing phonetics by computer’. 2022. [Online]. Available: http://www.praat.org/

[25] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, ‘Montreal Forced Aligner: Trainable text-speech alignment using Kaldi’, in *Interspeech 2017*, ISCA, Aug. 2017, pp. 498–502. doi: 10.21437/Interspeech.2017-1386.

[26] S. Gonzalez, J. Grama, and C. E. Travis, ‘Comparing the performance of forced aligners used in sociophonetic research’, *Linguist. Vanguard*, vol. 6, no. 1, p. 20190058, Jan. 2020, doi: 10.1515/lingvan-2019-0058.