Verification and Characterization of Users by Their Voices

Laura Fernández Gallardo

March 23, 2018

Abstract

Modern human-computer interaction systems may not only be based on interpreting natural language to determine dialog strategies but also on detected speakers' identity and their interpersonal characteristics. The goal of this work is to automatically characterize users from their speech signals, i.e. to recognize their identity and personality traits (confidence, friendliness, competence, etc.) by the sound of their voices and manner of speaking. In addition, this work evaluates the influence of different transmission channels, which degrade the quality of speech signals, on subjective and on automatic speaker recognition and characterization. Auditory tests have been conducted to assess significant effects of speech bandwidth on perceptive speaker ratings, and predictive models have been tested with speech degraded through a range of telephone transmissions involving different bandwidths, codecs, and other distortions. The multiple findings of these investigations may motivate the development of future personalization schemes based on the detection of speaker characteristics and exposed to speech quality degradations.

1 Introduction

1.1 Presentation

- In this presentation I am going through my work during the past few years. It focuses on the recognition of users' identity and of their social characteristics from their speech signals, that is, from the sound of their voices and the way they talk.

- Have you ever worked with speech signals? (If not) Well, I then hope that by the end of my presentation you get a nice overview of what we can do with speech signals, how we can work with them and what information we can get from the talker.

1.2 Who I am

This is just a little bit about myself:

- After I finished my Masters, I started my PhD on the topic of speaker recognition (recognition of users' identity based on voices). Collaboration between TU Berlin and University of Canberra, Australia. Funded by Deutsche Telekom.
- For my Postdoc: moved on to recognition of speakers' social characteristics. I wrote a project proposal and got DFG - Deutsche Forschungsgemeinschaft funding.

1.3 Global Outline

The general picture of my work can be depicted in a diagram like this:

- We have a person who is talking (the speaker)
- At the other end, we have a listener, who tries to recognize the speaker's identity. He is comparing the voice that he is listing to to the voices he knows from memory.
- We can of course perform this automatically, which is typical of biometric systems based on voice. The system's task is to detect whether the voice corresponds to the person the speaker claims to be, and it makes a decision to grant access to private information.
- In addition, humans can also perceive other speaker characteristics from the voice, such as speakers' personality. For example, whether the speaker is friendly, confident, mature, competent, and so on.
- And we can also perform this automatically, for example, in personalization applications. The goal here is not to recognize user's temporal emotions (like anger, happiness,...), but I rather focus on more stable speaker traits. You can see this as social attributes, or personality of talkers.

At this point, I also want to make clear that I am not concerned about recognizing the message or the words being spoken. In all cases, my focus is on the identity and personality characteristics, also conveyed in the voice sounds. However, neutral speech content needs to be controlled for not to bias the perception of users.

Between the speaker and the listener, or between the speakers and the automatic systems, there is a telephone communication. Different artifacts of the voice transmission path are going to affect the quality of the signal at the receiver's end. For instance, channel bandwidth, codec, and packet loss can have a big effect on the perceived speech quality.

This brings me to the main goal of my research: to evaluate the effects of telephone transmission on these four parts you see on the right.

- My PhD concentrated on human and automatic speaker recognition (the two parts on the top).
  - ⋆ The motivation for this project was to understand whether and how speaker-specific voice components are affected by which parameters of telephone distortions.
  - ⋆ One of my main contributions (we will see this later) was the demonstration of the benefits of enhanced, wideband channels, in contrast to narrowband communications - which motivates the deployment of this kind of network: Listeners were able to recognize voices more accurately and faster, and biometric speaker verification systems performed better when speech signals were transmitted through channels delivering good speech quality.
- In a similar way, in my Postdoc, I examined how the telephone distortions affect the speaker characterization performance, on the subjective and on the automatic side.
  - ⋆ For instance, if speakers are perceived less agreeable because of the channel degradations, this can directly affect the listeners' quality of experience of the communication. We want communication networks that permit the perception of users as with clean speech (no communication artifacts).
  - ⋆ Speaker characterization techniques can be very interesting for user personalization in human-computer communication systems. For example, the system can adapt the dialog strategy depending on the recognized users' personality. In this case, we do not want the performance to be affected by telephone distortions - I am evaluate whether this would happen.

I am going to stick to this outline throughout my presentation, and we are going to go through an overview of all parts. My intention is that you get a general idea of my work, so I will just present some representative experiments I conducted and their results without entering into much detail, and will leave that for the discussion part if there is interest from your side.

If you have any questions or would like more clarification please feel free to interrupt me at any point.

Let us then start by taking a deeper look into the kind of telephone degradations I am considering in my work.

1.4 Telephone Degradations

As you know, there are these bandwidths standardized in telephony. The difference is the range of voice frequencies that they can transmit.

We have the traditional narrowband (NB), which limits the speech signal to 300–3400 Hz. More recently, wideband (WB) channels have emerged, and they extend the upper limit to 7000 Hz (double), and the lower limit to 50 Hz: this has a large influence on perceived quality. An even more extended range of frequencies can be transmitted through super-wideband (SWB) channels, which deliver the best speech quality and are typically used for video-conferencing. WB and SWB can be found in VoIP, while NB is still predominant in the PSTN (public switched telephone network).

We can find human voice components up to 20 KHz or beyond, yet the range of human hearing is limited to 20 kHz, and to 20 Hz on the lower end.

The effect of bandwidth is illustrated in the following comparison.

- We can see the amplitude of the speech signal in the temporal domain and its spectrogram (frequency on the y axis). The lighter parts indicate higher energy regions.

- The signal on the top is a clean recording with 48 kHz sampling frequency - so, it has a bandwidth up to 24 kHz, and on the bottom we see the same speech degraded through a rather severe communication channel. In this example, the signal was transmitted through a narrowband bandwidth (that is why you do not see frequency components above 3400 Hz). The AMR-NB codec has been employed to compress and decompress the signal. Additionally, a random packet loss rate of 10% has been applied, this means that, in the transmission, 10% of the signal did not reach the receiver. And jitter of 10 ms was also applied, which means that there was network congestion and some packets arrive with delay. Here, we just treat all this as parameters affecting the signal quality.

Let us listen to these signals.

(...)

Now you have a feeling of how these distortions are affecting the speech quality.

All frequency components above 3400 Hz have been filtered out. A WB or SWB communication channel would have included more components that would have resulted in more natural speech and better intelligibility. The goal of my work is / has been to study how this influences speaker identification and characterization performances. There are no strong differences between listening to WB or to SWB speech, since human hearing presents greater resolution in the lower frequencies.

2 My PhD

As I introduced before, my PhD focuses on speaker recognition affected by telephone distortions [Fernández Gallardo, 2013].

2.1 Human Speaker Recognition

Regarding human speaker recognition, I will just briefly say that I conducted several listening tests with speech stimuli that were degraded through different telephone conditions.

The task for the participants of these tests was to indicate who has spoken, or whether two signals correspond to the same speaker or not.

Then, I examined the statistical significance of the results and: There is an improvement of WB over NB, but not of SWB. This might be due to the lack of speaker-discriminative frequency components added in SWB.

My publications on human speaker recognition influenced by telephone transmissions: [Fernández Gallardo et al., 2012a, Fernández Gallardo et al., 2013a, Fernández Gallardo et al., 2013b].

2.2 Automatic Speaker Verification

I think that automatic speaker verification (AVS) is more interesting for you.

2.2.1 Extracting Speech Features

First of all, I wanted to give you an idea of how speech is parametrized. This is a very simplified scheme for how to extract Mel-Frequency Cepstral Coefficients(MFCCs), very popular for speech and speaker recognition.

It is important to note that the recorded speech needs to be sliced into frames of about 20 ms–30 ms, and typically with overlap of 10 ms.

(The selection of the frame width is a tradeoff between temporal and spectral resolution -¿ we want the window to be long enough to average over local signal fluctuations, but short enough not to average over adjacent speech sounds.)

For each of the frames, a Fourier Transform is applied, and the mel-filterbank is applied to the power spectrum and then the coefficients are extracted. The Mel-scale aims to mimic the non-linear human ear perception of sounds (more resolution at lower frequencies). Typically, the first 10–12 MFCCs are retained.

2.2.2 ASV Evaluations

A typical biometric system for automatic speaker verification works like this:

The user makes an identity claim and needs to read a given text or to talk freely.

The system makes a comparison between the received speech signal and the model corresponding to the claimed identity, to decide whether it is the right speaker or an impostor. Of course, a model from each person has to be created beforehand from enrollment speech.

In my PhD, I employed the state-of-the-art GMM-UBM, JFA, and i-vector systems (2014). The current state-of-the-art employs i-vectors and DNN for speaker verification.

Before we go on, I would like to highlight the difficulty of training good speaker models with large datasets for my work. Unfortunately, there are not so many appropriate datasets that I can use. I need to start from clean - undistorted full-band speech and apply the different telephone distortions in a controlled manner. However, there are very few datasets with signals of sufficient sampling frequency. I need microphone speech with at least fs=32 kHz for simulating the SWB transmissions, yet most of databases contain NB degraded speech, since they were collected at the receiver end of a transmission. This has been the main limitation of my research.

Still, I could find several databases recorded in clean conditions and pooled many speech segments to perform different experiments:

One of the experiments with GMM-UBM employed a small dataset of clean speech sampled at 44.1 kHz. In this plot we can see the equal error rate (EER, the lower the better) across different degradation conditions. The EER is a performance metric, as you know, the value when false rejections and false acceptances are equal (this can be seen as a binary classification). I could show that there was an important EER reduction when we move from NB to WB, and from WB to SWB. This benefit was specially observed for female speech, since their voices have higher frequency components due to their shorter vocal tract compared to males.

My publications on automatic speaker recognition influenced by telephone transmissions: [Fernández Gallardo et al., 2012b, Fernández Gallardo et al., 2014d, Fernández Gallardo et al., 2014a, Fernández Gallardo et al., 2014b, Fernández Gallardo e

2.3 My PhD's Contributions

Here are the main contributions of my PhD, divided into human and automatic speaker recognition [Fernández Gallardo, 2016b (German reports: [Möller et al., 2015, Fernández Gallardo, 2016d]).

In both cases, there was an improvement in the transition to enhanced channels. Therefore, together with speech quality, speaker recognition can be considered as an additional criteria for the deployment of WB-capable networks and terminals [Möller et al., 2014, Fernández Gallardo and Möller, 2015, Fernández Gallardo et al., 2015]. SWB offered an improvement on the automatic side, which was not observed for human speaker recognition.

3 My Postdoc

When I started my Postdoc there was not suitable speech database available with clean speech and with the labels that I needed. So, I embarked myself into the task of designing and collecting a new database for my research.

3.1 New Appropriate Speech Database

300 speakers, that speak German as mother tongue, were recorded at our labs. The speakers were recorded in an acoustically-isolated room and we employed a high-quality microphone to record scripted and spontaneous speech. (I got great support from student workers who conducted the recording sessions). Then, I segmented and arranged all recorded files to prepare the release of this data to the scientific community. It is only available for research[1]. I got very good feedback and some researchers are already using this data resource [Fernández Gallardo, 2016c, Fernández Gallardo and Weiss, 2018].

I conducted a series of listening tests to collect labels for the database in terms of speaker characteristics. A semantic differential questionnaire was presented, with antonyms at both ends of a continuous scale. You can see the list of all 34 questionnaire items on the right (German).

The speech consisted on a dialog where speakers had to order a pizza (We listened to a start of one of these dialogs when I presented an example of telephone distortions). In this test, only clean speech was presented.

In total, 114 listeners participated to label the whole database (300 speakers). I considered the mean of their ratings as ground-truth: perceptions of speaker characteristics.

The 34-dimensional ratings were reduced by performing factor analysis to a smaller set of 5 dimensions (which I will call traits). These 5 traits are: warmth, attractiveness, confidence, compliance, maturity - and they can bee seen as perceptual dimensions of attributions that can be made by listening to speakers. Interestingly, the same names could be given to the male and to the female traits.

The distribution of speakers is shown in this pairplot (blue = male, orange = female - there are more females in the database). The warmth and attractiveness traits are correlated, as well as compliance and confidence (negatively), and confidence and maturity (makes sense).

I consider that warmth and attractiveness represent a space of positively and negatively perceived speakers. I have performed binary classification to discriminate between low and high WAAT speakers (we'll see this later).

[1]The ISLRN of this corpus is 157-037-166-491-1. The data has been made available at the CLARIN repository: `hdl.handle.net/11022/1009-0000-0007-C05F-6` under the CLARIN ACA+BY+NC+NORED license (freely available for scientific research).

3.2 Human Speaker Characterization

Let us go on by looking at how human perceptions of speakers can be affected by channel bandwidth.

In another round of listening tests (with different participants), I collected the same ratings by presenting speech in NB and in WB, from just a subset of 20 "extreme speakers". So, I got ratings to the 34 speaker characteristics in NB and in WB.

Then, another group of listeners provided ratings of speech quality to the same stimuli. The presented continuous scale ranged from "extremely bad" to "ideal".

With the data I got from these listening tests, I performed Spearman rank correlations between quality ratings and the ratings of every speaker characteristic, for males and for female separately (since they present different stereotypes). A strong correlation indicates that when a telephone channel provides better perceived speech quality, also higher ratings are given to that speaker characteristic.

For instance, we can see a strong correlation between perceived speech quality and ratings given to "likable" for males (blue) and, to a lesser extent, to "likable" for females (orange). The same can be observed for the ratings given to "pleasant". There are some gender differences: with higher quality, males are perceived as more compliant while females as more cynical. Similarly, with higher quality, males are perceived as more modest and adult while females as more impudent and childish. For other speaker characteristics, males and females tend to be perceived in the same direction with speech of higher quality.

In this graph, the characteristics are presented as in the semantic differential questionnaire, with antonyms at each side.

The speaker characteristics highlighted with channels providing better quality are:

- For males:

  - unobtrusive, characterful, sympathetic, likable
  - pleasant, attractive, beautiful
  - impersonal, competent, calm, adult, decided, secure

- For females:

  - friendly, sympathetic, characterful
  - interested, emotional, active, young
  - impudent, competent, secure, decided, intelligent

My publications on human speaker characterization: [Fernández Gallardo, 2016a, Zequeira Jiménez et al., 2017, Fernández Gallardo and Weiss, 2017b, Fernández Gallardo et al., 2017, Fernández Gallardo and Weiss, 2017a]. Influence of telephone transmissions: [Fernández Gallardo and Weiss, 2016, Fernández Gallardo et al., 2018, Fernández Gallardo, 2018].

4 My Postdoc: ongoing work

At the moment I focus my work on machine learning for speaker characterization.

4.1 Automatic Speaker Characterization

4.1.1 My Pipeline

Features: I work with the "eGeMAPS" feature set, which has been shown to provide good results for emotion recognition tasks. A set of 88 parameters that describe frequency, energy, spectral and temporal aspects can be extracted from each speech segment. For that I employ the OpenSMILE tool.

I used R and scikit-learn in python for data exploration and for building predictive models.

Reminder: the targets I consider in my experiments are the 34 item ratings, or the 5 trait scores derived by applying factor analysis. I am addressing classification and regression problems with these key performance metrics.

I use a pretty much standard pipeline for machine learning. I first split the data into train and test, and perform a nested hyperparameter tuning with the train set trying out different model families. Then, I evaluate the performance on the test set.

The nested hyper-parameter tuning consists on:

- The train data is split into a set A and a set B (80%–20%). The test set was already hold out before.
- For each model family, I perform a randomized or grid search of model hyperparameters with cross-validation.
- The best model is chosen based on their performance on set B
- Finally, the chosen model is trained with all trained data (A and B) and its performance is to be evaluated later on the test set.

4.1.2 Binary Classification (WAAT)

Here, I want to take a closer look at one of the relevant tasks I am addressing: the binary classification of high/low WAAT (warmth-attractiveness).

I applied K-Means to cluster the speakers into 3 classes: low - mid - and high WAAT. I then dropped all speakers that belong to the mid cluster to address binary classification.

Different classifiers were tuned and trained: Logistic Regression, Naive Bayes, K-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machines.

In a real application we would just select the classifier that gave the best performance on the development set (B). Here, I evaluated all classifiers on the test set.

I have plotted the average per-class accuracy over the different tuned classifiers. The blue line corresponds to the performance on the B set, and the orange line to the performance on the test set. Worst: Dummy, SVC poly. Best: SVC rbf, KNN.

I have several test speech segments for each speaker. I performed majority voting with the decision (high/low WAAT) made for each of the segments to come up with the final decision for the speaker. The following plots represent again the WAAT space, with each circle corresponding to one speaker of the test set. They are blue if the decision was correct or red otherwise, and their diameter indicate the strength of the decision of majority voting. For instance, if 60% of the speech segments were correctly classified and 40% were incorrectly classified, then the circle is blue (correct final decision), and smaller than others corresponding to speakers for which all decisions of their speech segments were correct.

Some mistakes have been made for speakers close to the mid WAAT region (red circles), for the SVC classifier (rbf kernel), with 80% accuracy. This can be compared to a dummy classifier based on random guessing. All circles are smaller (less confidence in the decision) and more speakers are miss-classified (accuracy was 50%).

4.1.3 Effects of Transmission Channels

Back to my pipeline: when I have performed experiments with degraded speech, I have trained my models with clean signals and evaluated the performance with degraded speech. This would represent an application where we have trained models with clean speech from existing datasets, and received degraded speech in production due to telephone transmissions. Again, the task is to classify speakers into high or low WAAT.

In this figure we can observe the average per-class accuracy given by the SVC classifier with rbf kernel (the best model with clean speech) across different transmission bandwidths and codecs (in this case, packet loss = 0 and jitter = 0). The points joined with a line correspond to the same codec, with different bit rates. The performance with clean speech was 72%. However, it drops to chance level with NB degradations. WB and SWB codecs tend to offer better performance, yet still far from that of clean speech. This seems to indicate that an application receiving telephone-transmitted speech would offer little performance compared to receiving clean speech. NB communications should be avoided.

Still, I would like to perform further experiments where the performance with clean speech is higher and re-evaluate the effects of communication channels. Particularly, I want to look into feature engineering and selection before model tuning and evaluation.

4.2 My Postdoc's Contributions

The contributions of my Postdoc are summarized in this slide for human and automatic speaker characterization. I have created a new valuable language resource, very much needed in this research field.

I have derived the main traits of speaker attributes that can be perceived from voices, and related speaker characteristics to speech quality and to voice descriptions (not seen in this presentation).

On the automatic side, I have explored the importance of features contributing to speaker characterization. I have written open-source code [2] with my pipeline to perform experiments in different configurations. Finally (still on-going) I am evaluating the speaker characterization performance given degraded test speech.

5 Summary

In this presentation we have seen the different parts of my research:

- The data I am working with: a new speech database labeled with speaker characteristics, released for the academic research community.
- Effects of telephone distortions on human and automatic speaker recognition. There is a significant difference in performance between NB and WB speech - the importance of frequency components beyond 3400 Hz is manifested and the deployment of WB communications is motivated.
- On the human speaker characterization side, speaker characteristics' perceptions have been explored via listening tests.
- Machine learning experiments show the influence of telephone distortions on the automatic speaker characterization performance.

References

[Fernández Gallardo, 2013] Fernández Gallardo, L. (2013). Speaker Recognition and Speaker Characterization over Landline, VoIP and Wireless Channels. In *Doctoral Consortium, International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 665–670.

[2] https://github.com/laufergall

[Fernández Gallardo, 2016a] Fernández Gallardo, L. (2016a). A Paired-Comparison Listening Test for Collecting Voice Likability Scores. In *Informationstechnische Gesellschaft im VDE (ITG) Conference on Speech Communication*, pages 185–189.

[Fernández Gallardo, 2016b] Fernández Gallardo, L. (2016b). *Human and Automatic Speaker Recognition over Telecommunication Channels*. T-Labs Series in Telecommunication Services. Springer-Verlag, Singapore, 1 edition. print/online.

[Fernández Gallardo, 2016c] Fernández Gallardo, L. (2016c). Recording a High-Quality German Speech Database for the Study of Speaker Personality and Likability. In *Phonetik und Phonologie im deutschprachigen Raum (PundP12)*, pages 43–46.

[Fernández Gallardo, 2016d] Fernández Gallardo, L. (2016d). Sprechererkennung - Auditive Wiedererkennbarkeit bei Breitband-Telefonie. *VDE-Dialog, ITG-News*, 1(1):17–18.

[Fernández Gallardo, 2018] Fernández Gallardo, L. (2018). Effects of Transmitted Speech Bandwidth on Subjective Assessments of Speaker Characteristics. In *International Conference on Quality of Multimedia Experience (QoMEX)*.

[Fernández Gallardo et al., 2018] Fernández Gallardo, L., Mittag, G., Möller, S., and Beerends, J. (2018). Variable Voice Likability Affecting Subjective Speech Quality Assessments. In *International Conference on Quality of Multimedia Experience (QoMEX)*.

[Fernández Gallardo and Möller, 2015] Fernández Gallardo, L. and Möller, S. (2015). Towards the Prediction of Human Speaker Identification Performance from Measured Speech Quality. In *Annual Conference of the International Speech Communication Association (Interspeech)*, pages 443–447. ISCA. print/online.

[Fernández Gallardo et al., 2012a] Fernández Gallardo, L., Möller, S., and Wagner, M. (2012a). Comparison of Human Speaker Identification of Known Voices Transmitted Through Narrowband and Wideband Communication Systems. In *Informationstechnische Gesellschaft im VDE (ITG) Conference on Speech Communication*, pages 219–222.

[Fernández Gallardo et al., 2013a] Fernández Gallardo, L., Möller, S., and Wagner, M. (2013a). Human Speaker Identification of Known Voices Transmitted Through Different User Interfaces and Transmission Channels. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7775–7779.

[Fernández Gallardo et al., 2015] Fernández Gallardo, L., Möller, S., and Wagner, M. (2015). Importance of Intelligible Phonemes for Human Speaker Recognition in Different Channel Bandwidths. In *Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1047–1051. ISCA.

[Fernández Gallardo et al., 2012b] Fernández Gallardo, L., Wagner, M., and Möller, S. (2012b). Analysis of Automatic Speaker Verification Performance over Different Narrowband and Wideband Telephone Channels. In *Australasian International Conference on Speech Science and Technology (SST)*, pages 157–160.

[Fernández Gallardo et al., 2013b] Fernández Gallardo, L., Wagner, M., and Möller, S. (2013b). Transmission Channel Effects on Human Speaker Identification in Multi-Party Conference Calls. In *International Conference on Information Technology in Asia (CITA)*, pages 38–43.

[Fernández Gallardo et al., 2014a] Fernández Gallardo, L., Wagner, M., and Möller, S. (2014a). Advantages of Wideband over Narrowband Channels for Speaker Verification Employing MFCCs and LFCCs. In *Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1115–1119.

[Fernández Gallardo et al., 2014b] Fernández Gallardo, L., Wagner, M., and Möller, S. (2014b). I-vector Speaker Verification based on Phonetic Information under Transmission Channel Effects. In *Annual Conference of the International Speech Communication Association (Interspeech)*, pages 696–700.

[Fernández Gallardo et al., 2014c] Fernández Gallardo, L., Wagner, M., and Möller, S. (2014c). I-vector Speaker Verification for Speech Degraded by Narrowband and Wideband Channels. In *Informationstechnische Gesellschaft im VDE (ITG) Conference on Speech Communication*.

[Fernández Gallardo et al., 2014d] Fernández Gallardo, L., Wagner, M., and Möller, S. (2014d). Spectral Subband Analysis of Speaker Verification Employing Narrowband and Wideband Speech. In *Odyssey 2014: The Speaker and Language Recognition Workshop*, pages 81–87.

[Fernández Gallardo and Weiss, 2016] Fernández Gallardo, L. and Weiss, B. (2016). Speech Likability and Personality-based Social Relations: A Round-Robin Analysis over Communication Channels. In *Annual Conference of the International Speech Communication Association (Interspeech)*, pages 903–907.

[Fernández Gallardo and Weiss, 2017a] Fernández Gallardo, L. and Weiss, B. (2017a). Perceived Interpersonal Speaker Attributes and their Acoustic Features. In *Phonetik und Phonologie im deutschprachigen Raum (PundP13)*, pages 61–64.

[Fernández Gallardo and Weiss, 2017b] Fernández Gallardo, L. and Weiss, B. (2017b). Towards Speaker Characterization: Identifying and Predicting Dimensions of Person Attribution. In *Annual Conference of the International Speech Communication Association (Interspeech)*, pages 904–908.

[Fernández Gallardo and Weiss, 2018] Fernández Gallardo, L. and Weiss, B. (2018). The Nautilus Speaker Characterization Corpus: Speech Recordings and Labels of Speaker Characteristics and Voice Descriptions. In *International Conference on Language Resources and Evaluation (LREC)*.

[Fernández Gallardo et al., 2017] Fernández Gallardo, L., Zequeira Jiménez, R., and Möller, S. (2017). Perceptual Ratings of Voice Likability Collected through In-Lab Listening Tests vs. Mobile-Based Crowdsourcing. In *Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2233–2237.

[Möller et al., 2015] Möller, S., Fernández Gallardo, L., and Wagner, M. (2015). Wiedererkennbarkeit von Sprechern bei schmal- und breitbandiger Telefonbertragung. In *Elektronische Sprachsignalverarbeitung (ESSV)*.

[Möller et al., 2014] Möller, S., Köster, F., Fernández Gallardo, L., and Wagner, M. (2014). Comparison of Transmission Quality Dimensions of Narrowband, Wideband, and Super-Wideband Speech Channels. In *International Conference on Signal Processing and Communication Systems (ICSPCS)*, page p41.

[Zequeira Jiménez et al., 2017] Zequeira Jiménez, R., Fernández Gallardo, L., and Möller, S. (2017). Scoring Voice Likability using Pair-Comparison: Laboratory vs. Crowdsourcing Approach. In *International Conference on Quality of Multimedia Experience (QoMEX)*.