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, 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 (detecting who is speaking). 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 DGF 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 emotions, 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 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.

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.

The main goal of my research is to evaluate the telephone transmission effects on the four parts you see on the right.

- My PhD concentrates 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 motivation for the deployment of enhanced, wideband channels, since these offer important benefits with respect to traditional communications.
  - ⋆ 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 this system.
  - ⋆ Speaker characterization techniques can be very insteresting for user personalitation 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 use 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. 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 this work.

1.4 Telephone Degradations

As you know, there are several 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. 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.

We can find human voice components up to 20 KHz or higher, 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 of the y axis). The lighter parts indicate higher energy regions. The signal on the top is a clean recording with 48 kHz sampling frequency, and on the bottom we see the same speech degraded through a very 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.

Do you have any question so far?

2 My PhD

As I introduced before, my PhD focuses on speaker recognition affected by telephone distortions.

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 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.

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 an example of 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 -¿ Window: 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 sound (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.

In my PhD, I employed the state-of-the-art GMM-UBM 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 models with large datasets for my work. I need to start from clean - undistorted 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 the SWB transmissions, yet most of databases contain NB speech and with degradations, since they were collected at the receiver end of a transmission. This has been the main limitation of my research.

One of the experiments with GMM-UBM employed a small dataset of clean speech sampled at 44.1 kHz. The preformance is reported in terms of equal error rate (EER, the lower the better), which is 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 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 compared to males.

2.3 My PhD's Contributions

Here are the main contributions of my PhD, divided into human and automatic speaker recognition.

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. SWB offered an improvement on the automatic side, which was not observed for human speaker recognition.

3 My Postdoc

I started my Postdoc project from just the motivation (Influence of telephone degradations on human and automatic speaker characterization).

There was not suitable speech data available, so I embarked myself into the task of designing and collecting a

new database for my research.

300 speakers, german as mother tongue, were recorded at our labs (a student worker conducted the recording sessions). The speakers were in an acoustically-isolated room and we employed a high-quality microphone to record scripted and spontaneous speech. Then, I segmented and prepared all recorded files to prepare the release of this data to the scientific community. It is only available for research.

3.1 Human Speaker Characterization

3.1.1 Labelling the NSC Corpus

I run a campaign to collect labels for the database in terms of speaker characteristics by performing many listening tests. 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.

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 to a smaller set by performing factor analysis. 5 dimensions (which I will call traits) were found for male and for female speech. In these tables you can see which questionnaire items were retained and their loadings on each of the traits. For instance, “sicher” and “unentschieden” load on the 3rd trait, and accordingly I named it “confidence”. The 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). Interestingly, the warmth and attractiveness traits are correlated, as well as conpliance and confidence, 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.

3.1.2 Effects of 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 ”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 scale ranged from “extrem schlecht” to “ideal”.

Here is the kind of data I got. There is a plot for males and another one for females. The y axis represents the mean opinion scores (MOS), which is just the mean of the ratings given to quality. On the x axis, there is “likability”, one of the 34 scales of the speaker characterics questionnaire. Each point corresponds to one speaker, and is color-coded by speech bandwidth. We can already see that, for males, there is some tendency of perceiving the same speaker as more likable if the stimuli was presented in WB instead of NB. This is a bit less clear for female speech.

From these data, I performed Spearman rank correlations between quality and ratings of every speaker characteristics. For males and for female separately (since they present different stereotypes). A strong correlation indicates that when a telephone channel provides better speech quality, also higher ratings are given to that speaker characteristic. In this graph, the characteristics are presented as in the semantic differential questionnaire, with antonyms at each side.

4 My Postdoc: ongoing work

4.1 Automatic Speaker Characterization

4.1.1 My Pipeline

Features: eGeMAPS (88) with OpenSMILE.

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. Here, I am only presenting binary classification of high/low WAAT. Key metrics for classification and regression.

The general pipeline of my experiments looks like this.

Nested hyper-parameter tuning.

4.1.2 Binary Classification (WAAT)

The WAAT space. 3 classes. Addressing high/low classificaiton.

Different classifiers tuned and trained.

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.

grouped by and averaged for speakers.

sorted from worse to best according to this performance (green line). Worst: Dummy, SVC poly. Best: SVC rbf, KNN.

speakers correctly and wronly classified in the WAAT space, for dummy (0.5) and for SVC rbf (0.8).

4.1.3 Effects of Transmission Channels

4.2 My Postdoc's Contributions

5 Summary

Reminder of motivation and applications of my projects.