On the Impact of Voice Encoding and Transmission on the Predictions of Speaker Warmth and Attractiveness

LAURA FERNÁNDEZ GALLARDO, Technische Universität Berlin, Germany
RAMON SANCHEZ-IBORRA, University of Murcia, Spain

Modern human-computer interaction systems may not only be based on interpreting natural language but also on detecting speaker interpersonal characteristics in order to determine dialog strategies. This may be of high interest in different fields such as telephone marketing or automatic voice-based interactive services. However, when such systems encounter signals transmitted over a communication network instead of clean speech, e.g., in call centers, the speaker characterization accuracy might be impaired by the degradations caused in the speech signal by the encoding and communication processes. This paper addresses a binary classification of high versus low warm–attractive (WAAT) speakers over different channel and encoding conditions. The ground truth is derived from ratings given to clean speech extracted from an extensive subjective test. Our results show that, under the considered conditions, the AMR-WB+ codec permits good levels of classification accuracy, comparable to the classification with clean, non-degraded speech. This is especially notable for the case of a Random Forest-based classifier, which presents the best performance among the set of evaluated algorithms. The impact of different packet loss rates have been examined, whereas jitter effects have been found to be negligible.

CCS Concepts: • **Networks** → Network reliability;

Additional Key Words and Phrases: speaker characteristics, transmission channels, predictive modeling, speech processing

ACM Reference Format:
Laura Fernández Gallardo and Ramon Sanchez-Iborra. 2019. On the Impact of Voice Encoding and Transmission on the Predictions of Speaker Warmth and Attractiveness. *ACM Trans. Knowl. Discov. Data.* 1, 1, Article 1 (January 2019), 17 pages. https://doi.org/10.1145/3332146

1 INTRODUCTION

The ability to automatically assess speakers' social and personality-related characteristics is desired in multiple novel systems that aim at offering individualized services. In this line, recent developments have led to speech assistants with excellent natural language understanding and synthesis capabilities [Masche and Le 2018]. However, the characterization of individuals and their intentions and behavior still needs to be improved in order to achieve even more human-like communications [Zukerman and Litman 2001].

During the last years, the recognition of users' social characteristics as predictor of their behavior have attracted the attention of both academia and industry with the aim of providing customized and contextualized services to end-users in different ways. First, depending on the detected user's social characteristics, such as dominance, the interlocutor may

Authors' addresses: Laura Fernández Gallardo, Technische Universität Berlin, Quality and Usability Lab, Berlin, 10587, Germany, laura.fernandezgallardo@tu-berlin.de; Ramon Sanchez-Iborra, University of Murcia, Information and Communication Engineering Department, Murcia, 30100, Spain, ramonsanchez@um.es.

Manuscript submitted to ACM

adapt his/her speaking characteristics in order to make the user feel more comfortable during the conversation. Besides, in case of a commercial transaction, the vendor can detect the possible interest of the customer on the offered products.

A set of speaker interpersonal factors has been identified in [Fernández Gallardo and Weiss 2018] as prominent dimensions that can be detected from voices of unknown individuals. Concretely, these factors are: *warmth*, *attractiveness*, *confidence*, *compliance*, and *maturity*. The automatic detection of these user traits is also possible from measurable acoustic speech parameters such as frequency- and spectral-related features, as examined in [Fernández-Gallardo and Weiss 2017]. The speech data and accompanying labels of [Fernández Gallardo and Weiss 2018] constitute the newly released Nautilus Speaker Characterization (NSC) corpus, which we employ in this work. In particular, we approach a binary classification of speakers' warmth–attractiveness (WAAT) using as speech material customer-inquiry dialogues in German language.

In particular, we examine how the speaker WAAT classification performance in clean speech is affected by degradations inserted by bandwidth limitation, speech coding and network-introduced distortions. The two classes considered as ground truth have been determined by binarizing factor scores derived from extensive subjective tests, where clean, undistorted speech was presented to a group of raters. Therefore, this study evaluates whether the performance obtained with transmitted speech is comparable to that reached with clean speech, despite the channel degradations.

The most dominant interpersonal factor found in [Fernández Gallardo and Weiss 2018], *warmth*, coincides with the *benevolent* dimension of the interpersonal circumplex, playing a crucial role in the formed interpersonal impressions of speakers [Jacobs and Scholl 2005; Wiggins et al. 1988]. Besides, *attractiveness*, the second most relevant dimension, has its foundation on interpersonal attraction [Aronson et al. 2009]. The subjective perception of this trait is prone to be affected by degradations of telecommunication channels, together with the perception of voice quality, as shown in [Fernández Gallardo 2018].

The focus of this work is on assessing how two well-known classification algorithms, namely, Random Forest classifier (RF) and Vector Machines for classification with radial basis function (SVC-rbf), perform in the WAAT classification task with clean voice signals impaired by degradations and distortions. We intend that the presented outcomes may be valuable for different kind of telecommunication-based services, e.g., call centers, telemarketing, customer services, etc., where it is of relevant importance to classify speakers' personality based on characteristics of their transmitted voice.

The transmission channels considered in this work vary in frequency bandwidth: narrowband (NB, 300–3,400 Hz), wideband (WB, 50–7,000 Hz), and super-wideband (SWB, 50–14,000 Hz). While conventional NB is still predominant in the public switched telephone network (PSTN), today's VoIP services can offer WB and SWB speech, resulting in a better clarity and a more natural conversation. We also employ in this study a variety of codecs at different bit-rates, which introduce nonlinear distortions in the speech signal. Besides, by means of a controlled test-bench, realistic network conditions have been emulated by introducing jitter and packet loss to the transmission stream. Thus, the end-to-end communication channel is explored and the effects introduced in different steps, namely, encoding and transmission, are identified. Therefore, regarding the prediction accuracy of speaker WAAT, the main contributions of this work are the following: (i) showing possible advantages of employing enhanced bandwidths for voice communications, (ii) evaluating the influence of different coding schemes, and (iii) assessing network transmission effects (packet loss and jitter). To the best of authors' knowledge, none of previous works in this field has addressed the impact of both coding and network-introduced distortions on the classification of speaker's personality attributes such as WAAT.

The rest of the paper is organized as follows. Section 2 reviews relevant work regarding the automatic recognition of speaker characteristics. The speech database employed in this study is described in Section 3, along with the test-benches used for voice encoding and transmission. Section 4 details the speaker classification pipeline employed in our machine

learning-based experiments. Section 5 presents and discusses the attained results. Finally, the paper ends by summarizing the most important findings in Section 6.

2 RELATED WORK

A rapidly-growing interest in human-computer interactions and spoken dialog systems has been observed in the last couple of years. Chatbots are becoming increasingly predominant, especially for customer service and personal companions and assistants [Sarikaya 2017]. Main efforts are being undertaken on automatic speech recognition (ASR) and natural language understanding (NLU), facing challenges such as background noise in rooms, overlapping speech, and understanding context [Wang and Yuan 2016].

With the need of providing personalized, tailored solutions based on users' individual behavior and preferences, adaptive voice-based interactions are todays' focus of numerous applications in academia and in industry [Argal et al. 2018; Bellegarda 2014; Spillane et al. 2017]. Adaptive human-machine spoken dialog systems are already able to react to changes in users' emotions or attitudes. Techniques of emotion recognition are being adopted to achieve this adaptation, often combining different data sources such as facial expression, body movements, physiological signals, speech, etc. Thus, these systems can dynamically respond to affective speech by generating different dialog strategies [Litman and Forbes-Riley 2014] or by synthesizing emotions [Aly and Tapus 2013]. The field of emotion recognition is currently considering the most recent advances in deep learning, e.g. transfer learning [Gideon et al. 2017], multi-task learning [Parthasarathy and Busso 2017], and bag-of-audio-words [Schmitt et al. 2016], inspired by text-mining applications.

Besides emotional status, speakers' interpersonal characteristics are also strong indicators of users' preferences and individual behavior and intentions, which could provide additional cues for dialog adaptation. Different efforts have been conducted in the last years to detect speakers' big five personality traits [Mohammadi and Vinciarelli 2012] and their likability [Burkhardt et al. 2011], charisma, or persuasiveness, among others. Linguistic voice-quality parameters linked to speakers' attributes have been examined in a recent work in cross-language perception tests [Ní Chasaide et al. 2017]. Concurrently, crowdsourcing is emerging as the preferred strategy for efficiently labeling large datasets with subjective perceptions of speakers [Burmania and Busso 2017]. Efforts are still needed in order to guarantee similar label quality between crowdsourcing and laboratory-controlled listening tests [Eskénazi et al. 2013].

The interpersonal circumplex, which is the basis of manifold studies on human social relations, is defined by the two orthogonal dimensions *Competence/Dominance* and *Warmth/Benevolence*. These user traits can be crucial to determine their attitudes in dialog interactions. For instance, the classification of overlapping speech as competitive or non-competitive has been explored in [Chowdhury and Riccardi 2017] by employing acoustic and lexical features, and the authors of [Silber-Varod et al. 2017] performed speech feature selection to discriminate between leader and follower roles.

The different speaker attributes reviewed are subject to modifications when voices are transmitted through communication channels. These channels introduce artifacts through degradations and distortions inherent to the transmission process, e.g. speech coding. Some channels (e.g. wideband) provoke less alteration of voice properties than others (e.g. narrowband). It has been shown that, as the speech transmission bandwidth is extended, speech quality can be estimated to be significantly higher [Möller et al. 2006], whereas automatic speaker verification [Fernández Gallardo 2016], speech intelligibility [Teng and Kubichek 2006], and automatic speech recognition [Fernández Gallardo et al. 2017] performance is greatly improved, compared to channels limiting the speech bandwidth.

A typical example of application confronted with this kind of degraded speech would be an interactive response systems that incorporates the prediction of users' characteristics in call centers. In this scenario, the interest may be on detecting the emotions of users in order to determine their satisfaction with the provided service. Another example would be telemarketing applications, where the transmitted voice might exhibit different characteristics depending on the particularities of the channel transmission. The effects of different speech degradations on emotion recognition have been explored in recent studies. A modulation filtering method has been proposed in [Pohjalainen and Alku 2014] to overcome noisy telephone speech, whereas the effects of background noises and of low-pass filtering have been explored in [Vasquez-Correa et al. 2015] and in [Frühholz et al. 2016], respectively. GMM-based classification has been employed in [García et al. 2015] and in [Albahri et al. 2016] for emotion classification tasks where speech segments were degraded by the AMR codec in different bandwidths.

Despite it is generally assumed that communication degradations and distortions affect voice quality, very little attention is generally paid to the particularities of the data recordings employed for predictive modeling of speaker characteristics. Apart from the mentioned contributions, no investigation is known to the authors that examined the influence of transmitted speech on speaker stable, non-emotional characteristics such as those addressed in this paper. The parameters of the communication channels in telephone datasets are treated as a "black box", and no comparison between different transmission conditions has been found already published. Thus, in this work a comprehensive study evaluating the impact of both encoding and transmission processes on the automatic prediction of speaker warmth and attractiveness is presented. We stress the use of an important variety of codecs, bit-rates, and network conditions, with the aim of determining the effects introduced on clean speech.

3 SPEECH DEGRADATIONS AND DISTORTIONS

3.1 Microphone Speech Segments

We have employed the NSC corpus for this work [Fernández Gallardo and Weiss 2018]. Taking advantage of its sampling frequency of 48 kHz, the influence of the newly deployed SWB channels can be also studied together with the NB and WB effects (NB, WB, and SWB signals have 8, 16, and 32 kHz sampling frequency, respectively). Other databases with emotion or personality labels present already transmitted speech or 8 kHz sampling frequency [Burkhardt et al. 2011; Mohammadi and Vinciarelli 2012], and are therefore not suitable for this study.

The speech data consists on semi-spontaneous dialog turns from 300 German speakers (mean duration=23.0 s, SD=3.3s), that consisted on ordering a pizza. This conversation is representative of customer inquiries dialogues, which can be generalized as a typical commercial telephone service. As stated above, this kind of transactions are notably influenced by the perception that the end-user has about his/her interlocutor. Aiming at easing the reproducibility of our experiment to the rest of the research community, this speech database has been made publicly available[1]. The speech has been labeled by an average of 15 (range: 12–18) external raters on 34 interpersonal speaker characteristics, employing continuous numeric semantic-differential scales. As detailed in [Fernández Gallardo and Weiss 2018], factor analyses have been conducted and the 34 dimensions were reduced to the five perceptual factors already mentioned: *warmth*, *attractiveness*, *confidence*, *compliance*, and *maturity*.

K-means clustering has then been applied to the factor scores of the first two dimensions, *warmth* and *attractiveness* (WAAT) of the 300 speakers. The scores given to these two factors are correlated (r=.74 and r=.79, for male and for female speakers respectively). This indicates that speakers perceived as warm are also generally perceived as attractive,

[1] https://www.qu.tu-berlin.de/?id=nsc-corpus

Table 1. Codecs Under Consideration

Codec	Bandwidth	Bit-rate (kbps)
G.711 (A-law)	NB	64
G.711 (u-law)	NB	64
G.723.1	NB	5.3, 6.3
GSM-EFR	NB	12.2
AMR-NB	NB	12.2 4.75, 5.15, 5.9, 6.7, 7.4, 7.95, 10.2, 12.2
Speex-NB	NB	2.15, 11, 24.6
G.722	WB	64
AMR-WB	WB	6.6, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05, 23.85
AMR-WB+	WB	10.4, 12, 13.6, 15.2, 16.8, 19.2, 20.8, 24
G.722.1c	SWB	24, 32, 48
EVS	SWB	7.2, 24.4, 32, 48, 64, 96, 128
OPUS	SWB	24, 32, 48, 64, 128, 160

and vice versa. Since the perceptual ratings of speaker characteristics tend to be normally distributed [Weiss and Burkhardt 2012], k=3 has been chosen to identify three classes: low, mid, and high WAAT.

We address the binary classification problem, considering only the speakers of the low and the high class. Speakers from the mid class are hence excluded. We decided to focus on the automatic detection of those speaker classes for which higher agreement among labellers is assumed. Speakers belonging to the mid class are deemed to be 'difficult' to classify even for human annotators, and were hence excluded from our experiment. The high and low WAAT classes can be regarded as somewhat balanced: there are 34 and 46 male speakers in the low and in the high class, respectively; and 45 and 58 female speakers in the low and in the high class, respectively.

Short utterances from the 183 speakers are to be classified into high or low WAAT. The recorded semi-spontaneous dialog turns corresponding to each speaker were chunked into smaller segments, in order to have enough speech samples for the classification task. The files employed for training our classifiers correspond to four semi-spontaneous dialog turns: 'car rental booking', 'pizza order', 'book from the library', and 'doctor's appointment' [Fernández Gallardo and Weiss 2018], each chunked into three segments of approximately 20 ms. In all, the training set comprised 2196 utterances.

For testing our classification models, we employed NSC stimuli files (modified shortened versions of the 'pizza order' dialog) each chunked into two segments of about 10 ms. This resulted on a total of 366 utterances with different content than the utterances assigned to model training.

Communication channel degradations have been systematically applied to the clean test speech via voice-encoding processes and its subsequent transmission over an emulated network. Next subsections describe the computer-based test-benches deployed for voice encoding and transmission, respectively.

3.2 Voice Encoding

Different versions of the speech data have been created, and used for the speaker binary classification task. First, the speech signals were downsampled to 8, 16, or 32 kHz for NB, WB, and SWB transmissions, respectively, using anti-aliasing low-pass FIR filters. The speech was then level-equalized to -26 dBov, a characteristic level of telephone channels,

using the voltmeter algorithm of the International Telecommunication Union - Telecommunication Standardization Sector (ITU-T) Recommendation P.56 [ITU-T Recommendation P.56 1993]. Afterwards, the signals were processed with a pass-band filter simulating typical terminals' sending characteristics, as specified in [ETSI TR 103 138 2016], which are different for each bandwidth. Then, pass-band filters with the channel bandwidth-limiting characteristic were applied, following ITU-T recommendations:

- For NB channels: The 8-kHz P.48 Intermediate Reference System (IRS) weighting filter [ITU-T Recommendation P.48 1988] was applied as pre-processing. This filter is referred to as 'IRS8' in ITU-T Rec. G.191 [ITU-T Recommendation G.191 2000]. The bandwith filter applied was the ITU-T Rec. G.712 filter (300–3,400 Hz).
- For WB channels: A "modified P.341" band-pass filter with cut-off frequencies of 135–7000 Hz was applied as pre-processing to simulate the terminal and the bandwidth limitation. The low cut-off frequency of 135 Hz roughly represents a real WB terminal (mobile phone of reasonable quality or a good quality hand-held or hands-free device). Such a filter was also employed in [ETSI EG 202 396-2 2006] (Chapter 7.4).
- For SWB channels: The flat 50–14000 Hz filter was applied to simulate a good SWB-capable device and the SWB channel bandwidth. This filter can be found in ITU-T Rec. G.191 [ITU-T Recommendation G.191 2000] named as '14KBP'.

After band-filtering the signals, codecs were applied employing standard ITU or European Telecommunications Standards Institute (ETSI) tools or the open-source Speex codec. Finally, the speech was again level-equalized to -26 dBov.

The codecs and bit-rates (kbps) under consideration are shown in Table 1.

3.3 Voice Transmission

Once the speech signals were pre-processed for introducing the effects of enconding and terminals, they were transmitted over an emulated network, which was kept under perfect control. We adopted this approach in order to avoid undesired effects introduced by the networking system. Therefore, the transmission, delivery, and reception processes of the speech-streams were performed over an unique computer by means of virtualization techniques. Concretely, three Linux-based (Ubuntu 16.04 LTS) instances were launched in order to configure the architecture shown in Fig. 1. Both ends of the communication, namely, transmitter (TX) and receiver (RX), made use of the *ffmpeg* tool [ffmpeg 2018] for the streaming process. Please, note that no transcoding techniques were used, hence this tool was employed just for packetization and transmission purposes, keeping the codec employed in the previous steps. Each transmission was performed in real time, i.e., each packetized transmission had the the same duration than the contained speech signal. In order to precisely emulate the impairments introduced by a typical packet-switched network, a third instance similar to those described above was also launched. In this case, we made use of the Linux-based *traffic control* (tc) tool, which permits to configure the kernel packet scheduler. Thus, the transmitted streams from TX went through this virtual machine, which introduced the corresponding levels of packet loss and jitter, and forwarded the altered stream to RX (please, see Fig. 1).We have considered packet loss and jitter as the network-introduced degradations due to they are the most relevant impairments introduced by current packet-switched networks [Holub et al. 2018]. For these experiments, four levels of packet loss rate (PLR): 1%, 3%, 5%, and 10%, and two levels of jitter: 0 ms and 10 ms were considered. Both impairments were artificially introduced following a bursty distribution, typical in IP-based networks. Neither error correction techniques nor jitter buffer were employed in the receiver in order to evaluate the pure impact of the transmission process.

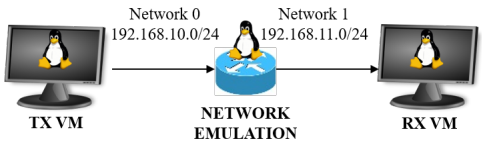

Fig. 1. Deployed test-bench.

Table 2. Average per-class accuracy of the considered classifiers

Classifier	Accuracy
Dummy	0,50
K-Neighbors	0,62
Logistic Regression	0,66
SVC-rbf	0,71
RF	0,76

4 SPEAKER CLASSIFICATION

4.1 Speech Features

As speech features we considered the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS). This set of 88 features, recommended by a group of experts in paralinguistics, can lead to satisfactory performance in the emotion recognition task, and also comparable to other larger (6373 features) brute-force parameter sets [Eyben et al. 2016]. As specified in [Eyben et al. 2016], frequency, energy, spectral, and temporal-related parameters are extracted from Gaussian-windowed speech frames of 60 ms long extracted every 10 ms. After smoothing over time with a symmetric 3-frame moving average, functionals such as mean, coefficient of variation, and percentiles are computed and included in the final set of 88 eGeMAPS features.

The openSMILE toolkit [Eyben et al. 2013] was employed to extract this speech feature set from the speech utterances of train and test sets described in Subsection 3.1.

4.2 Predictive Modeling Pipeline

Different binary classifiers have been built with clean speech segments (the training set) and tested with clean or transmitted versions of the test set. This scenario mimics the typical situation in which a telecommunication provider has at disposal a set of microphone-speech samples for training its algorithms but the network-introduced artifacts encountered in test cannot be defined beforehand due to the random nature of the IP-based communication networks.

As stated above, two model families have been chosen for training, based on their performance in prior preliminary experiments involving clean speech exclusively, namely, support vector machines for classification (SVC) with radial basis function (rbf) kernel, and random forest classification (RF). Other classifiers based on k-neighbors and logistic regression procedures were also initially considered but they were discarded from the present study due to their poor performance in preliminary tests in comparison with those finally selected. Table 2 presents the results, in terms of average per-class accuracy (per speaker), attained in these preliminary tests, which evaluated the performance of each classifier when discriminating speaker's WAAT under clean-speech conditions. Observe that these outcomes lead us to finally choose the classifiers based on SVC-rbf and RF techniques (both of them over 70% of accuracy).

The hyper-parameter tuning approach adopted for model building consisted on a) splitting the training data into both set A and set B, containing the 80% and 20% of the utterances, respectively, and b) randomized search of best speech features and of model's hyper-parameters with 10-fold cross-validation. Feature selection was performed based on the four highest label/feature ANOVA F-values. The average per-class accuracy was considered as a performance metric.

For tuning SVC-rbf classification, the cost paremeter (C) ranged 5 steps spaced evenly on a log scale from 1e2 to 1e4; and the values tried for rbf's gamma were 1e-5, 1e-4, and 1e-3. The rest of hyper-parameters were left to their default configuration in the 'scikit-learn' implementation. After tuning employing cross-validation on set A, the resulting the best-performing parameters were cost C= 3e3, rbf's gamma = 1e-3. 77 out of the 88 speech features were selected.

For RF classification, we limited the search for the optimum number of trees to the range 4–100 (random search) and narrowed down in a following step to the range 27–34 trees. The 'gini' criterion was used for splitting and no maximum depth of the tree was tuned. The rest of hyper-parameters were left to their default values in the 'scikit-learn' implementation. 75 out of the 88 speech features were selected and the number of trees leading to the best performance with cross-validation on set A was 33.

5 RESULTS

In this section we provide insights about the performance of the classification models under consideration, namely, RF and SVC-rbf. We focus on the effects of the impairments introduced by the encoding and transmission processes on the binary classification performance. Recall that the objective of these classifiers is the binary classification of speakers into high or low warmth-attractiveness (WAAT). For the sake of clarity, the effects of both voice encoding and network transmission are discussed independently with the aim of understanding the impact of each one on the accuracy of the decision algorithms.

Throughout the presentation of the results, we consider the *"per speaker"* average per-class classification accuracy. This metric combines the performance of classifying speakers as belonging to the low WAAT class with that of classifying speakers as belonging to the high WAAT class. It was computed as follows. Given the predictions of WAAT made for each of the test speech utterances (these were a total of 366 utterances), we performed majority voting with the utterances that had been produced by the same speaker (please note that each speaker produced two utterances). From the resulting decisions for the 183 speakers of this study, we then computed the average of the accuracy for each class (high or low WAAT), i.e. the macro-averaged recall score. This will be referred to as "accuracy" in the remainder of this paper.

5.1 Effects of Voice Encoding

As described previously, a relevant set of narrowband, wideband, and super-wideband codecs operating at different bit-rates has been considered in this study (Table 1). The accuracies considering the effects of the codec schemes and their associated bit-rates are presented and discussed in the following. In this case, the impairments introduced by the network, i.e., packet loss and jitter were set to zero. Fig. 2 and Fig. 3 depict the performance of Random Forest (RF) and of the Support Vector Machines (SVC-rbf) classifiers, respectively. The accuracies attained when using clean speech data are indicated by the dashed line. These were 0.76 and 0.71 for RF and for SCV-rbf classification, respectively.

Although in general the RF-based algorithm presents greater level of accuracy in comparison with the SVC-rbf one, similar tendencies regarding the performance attained for each codec can be identified. Firstly, the use of NB codecs (on the left in both figures) leads to a poor performance of both classifiers, in some cases even below the chance level (0.50).

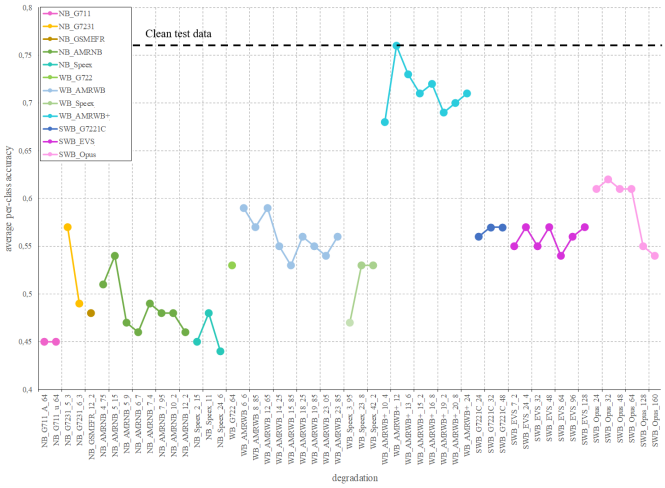

Fig. 2. Performance of the RF classifier with coded-decoded speech.

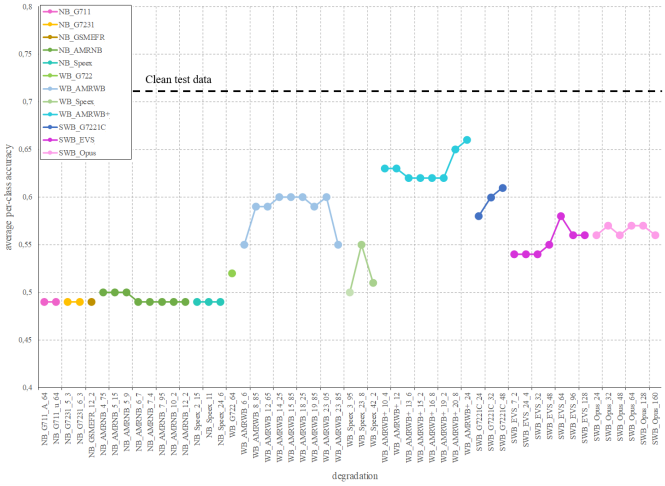

Fig. 3. Performance of the SVC-rbf classifier with coded-decoded speech.

Only the RF test performance with G.723.1 at 5.3 kbps has been found to be somewhat higher (0.57) than for the rest of NB codecs. The low accuracy of NB codecs is not related to the bit-rate employed since, for example, both G.711 codecs, which make use of the greatest bit-rate among the NB codecs, i.e., 64 kbps, lead the poorest performance. In addition, by considering the same codec with multiple bit-rates, e.g., AMR-NB, it cannot be determined that increasing or decreasing the bit-rate may lead to a better performance. Thus, it can be concluded that narrowing the encoding bandwidth has a remarkable negative impact on the automatic WAAT classification, but the bit-rate does not seem to have a notable effect in this task.

It can be seen that WB and SWB codecs permit a similar improvement on the classifiers performance with respect to the NB alternatives. Regarding WB codecs, the performance with G.722 and with Speex-WB is comparable to that of

Table 3. Highest average per-class accuracy for each codec

Codec	RF		SVC-rbf	
	Bit-rate (kbps)	Avg. per-class accuracy	Bit-rate (kbps)	Avg. per-class accuracy
G.711 (A-law)	64	0.45	64	0.49
G.711 (u-law)	64	0.45	64	0.49
G.723.1	5.3	0.57	5.3	0.49
GSM-EFR	12.2	0.48	12.2	0.49
AMR-NB	5.15	0.54	5.15	0.5
Speex-NB	11	0.48	11	0.49
G.722	64	0.53	64	0.52
AMR-WB	12.65	0.59	14.25	0.6
Speex-WB	23.8	0.53	23.8	0.55
AMR-WB+	12	0.76	24	0.66
G.722.1c	32	0.57	48	0.61
EVS	24.4	0.57	64	0.58
OPUS	32	0.62	32	0.57

NB codecs, although this performance is surpassed by the AMR-WB codec. It is worthwhile to mention the case of the AMR-WB+ codec, which clearly leads to the best performance of both classifiers. Focusing on the RF classification, many bit-rates employed by this codec allow notable accuracy levels over 0.70 with special mention to the case of 12 kbps that permits an accuracy level of 0.76, which is the same level than that attained with clean speech. The performance of SWB codecs appears to be comparable to that offered by the AMR-WB codec. While the performance with G.722.1C is prominent for SVC classification in SWB (Fig. 3), the Opus codec (at a bit-rate of 64 kbps or lower) performs better than other SWB codecs with RF classification (Fig. 2). In order to ease the readability of the afore-discussed outcomes, Table 3 presents the best result for each codec in terms of accuracy of both classifiers under consideration. Concretely, it shows the most accurate result and the associated bit-rate for each of the studied codecs.

Although, at first, one may think that the use of SWB codecs should lead to a better performance of the classification tasks, the attained results do not show such behavior. It should be noted that SWB speech codecs permit the transmission of the high-frequency range 7,000–14,000 Hz in comparison to WB. In the lights of these results, it might be inferred that these high-frequency components do not transport relevant information for our classification task. The WB frequency range (50–7,000 Hz) seems to be sufficient to convey the needed speech cues.

In order to provide statistical evidence about the afore-discussed results, the Binomial statistical test[2] has been also conducted in order to test for statistically significant differences in classification accuracy across different channel degradations, for both classifiers. The null hypothesis is that "the probability of correct classification with two different speech codecs is the same" [Noirhomme et al. 2014]. The Binomial test was performed with the classification results from each pair of degradations (coded-decoded speech). Significance level was set to $p<0.01$.

According to the results from the performed tests, the following remarks can be made: i) For both classifiers, performance significantly over chance level is found for clean speech and AMR-WB+ codec (all bit-rates); ii) the accuracy with clean speech is statistically significantly higher than that with all NB, WB, and SWB codecs, except for the case of AMR-WB+; iii) focusing on the RF classification, the AMR-WB+ codec offers statistically significantly superior accuracy than the rest of codecs in all bandwidths; iv) no notable differences between the rest of WB and NB

[2] https://github.com/laufergall/ML_Speaker_Characteristics/blob/master/classification/significance_tests_channels.md

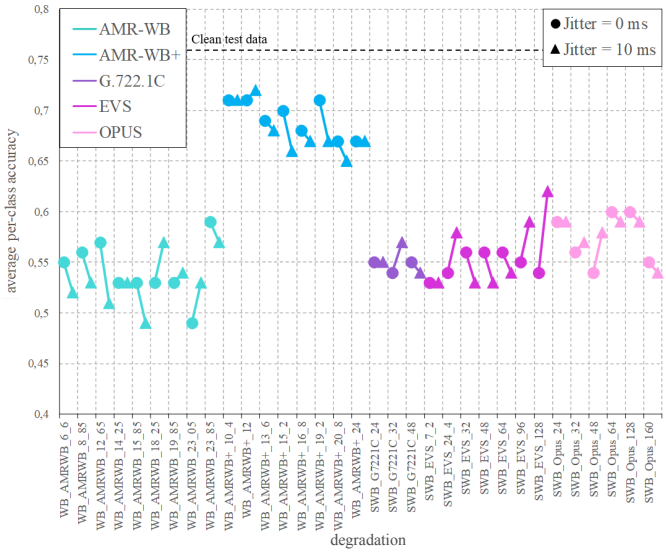

Fig. 4. Average per-class accuracy of RF classifier with network-introduced impairments. PLR = 1%.

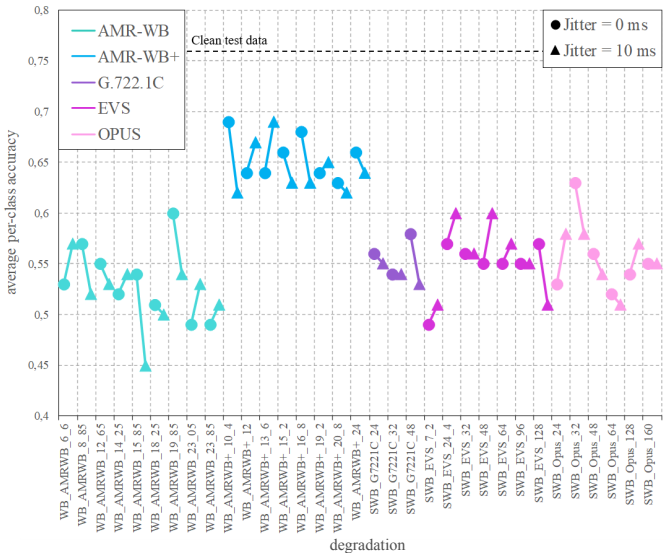

Fig. 5. Average per-class accuracy of RF classifier with network-introduced impairments. PLR = 10%.

codecs are found, except for AMR-WB, which offers higher performance than G.711 (in its two versions), in this line, the three SWB codecs also improve the performance of G.711 significantly; v) for the SVC-rbf classifier, the AMR-WB+ codec offers statistically significantly superior accuracy than the rest of codecs in all bandwidths, except for the case of AMR-WB, where no statistical differences between both codecs is detected; and vi) in general, all WB and SWB

codecs offer statistically significantly superior accuracy than the NB codecs, except for G.722 and Opus evidenced by no statistical differences in accuracy between G.722 (WB) and NB codecs, and between Opus (SWB) and AMR-NB.

5.2 Effects of Voice Transmission

In the following, we explore the impact of the impairments introduced by the network on both WAAT classifiers. In order to simplify this analysis, we discard those codecs that lead to low level of accuracy as seen in the previous section. Although not presented here, the behavior of the classifiers with these codecs remains notably poor when transmission-related distortions are also considered. Concretely, all the narrowband codecs and the G.722 and Speex wideband codecs have been discarded for the following analysis.

Fig. 4-5 and Fig. 6-7 present the performance of the RF and SVC-rbf classifiers, respectively, for a set of WB and SWB codecs and their corresponding bit-rates when the transmissions were affected by jitter and packet loss events. As stated above, we have a introduced controlled jitter of 0 ms and 10 ms, and a packet loss rates (PLR) of 1%, 3%, 5%, and 10%. In these figures we only present the performance of the classifiers with the extreme values of the PLR range, namely, 1% and 10%.

From a general perspective, and focusing on the impact of jitter, it can be seen that there is no clear decrease on the classifiers' accuracy with the appearance of jitter (10 ms) in any of the studied conditions. Regardless of the employed codec or bit-rate, the introduction of a notable level of jitter on the communication channel does not affect to the performance of the classifiers, evidenced by non-significant differences between the accuracy obtained with or without jitter. This performance is justified by the fact that jitter events do not introduce loss of information in the voice stream; instead, a packet reordering is produced due to the absence of de-jitter buffer, hence the speech fragments are received in an unexpected order but all the speech cues are kept.

However, focusing now on the other network-introduced impairment, i.e., packet loss, it can be seen that high PLR such as 10%, leads to generally lower levels of accuracy for both classifiers (comparing Fig. 4 with Fig. 5 and Fig. 6 with Fig. 7). This behavior is more notable in the case of the RF classifier, which suffers a greater decay in its overall performance in comparison with the SVC-rbf algorithm. Thus, the latter presents a greater robustness against the effect of packet loss; even with some specific codecs (e.g., AMR-WB), its accuracy is higher than that of the RF classifier.

Therefore, comparing these results with those attained with degraded (not network-transmitted) speech (Fig. 2 and Fig. 3), it can be concluded that RF performs better under ideal conditions (no artifacts in voice transmissions). The best overall performance is still attained by this classifier when using the AMR-WB+ codec. However, it presents some difficulties when processing network-transmitted signals.

Regarding the performance of the classifying algorithms with each particular codec, again the AMR-WB+ permits the best results for both classifiers, with a superior accuracy of the RF one. It should be observed that depending on the coding-rate employed, the RF classifier achieves accuracy levels over 0.70 and 0.65 with PLR = 1% and PLR = 10%, respectively. These figures are notably higher than the accuracy obtained with other encoding schemes as in the case of using clean speech (Fig. 4 and Fig. 5). Similar behavior is attained with the SVC-rbf classifier, although the accuracy reached by using the different codecs and bit-rates presents more homogeneity (Fig. 6 and Fig. 7). Therefore, it can be concluded that the AMR-WB+ codec is the most appropriate one among the big set of considered codecs for the automatic WAAT detection by means of the studied classifiers.

We now comprehensively study the effects of PLR on the classification performance with the AMR-WB+ codec specifically. Fig. 8 and Fig. 9 show the accuracy of both classifiers with an increasing level of PLR affecting the

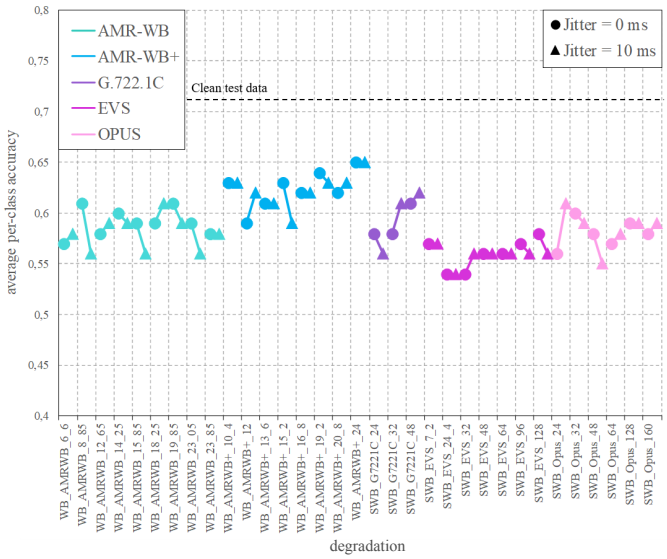

Fig. 6. Average per-class accuracy of SVC classifier with network-introduced impairments. PLR = 1%.

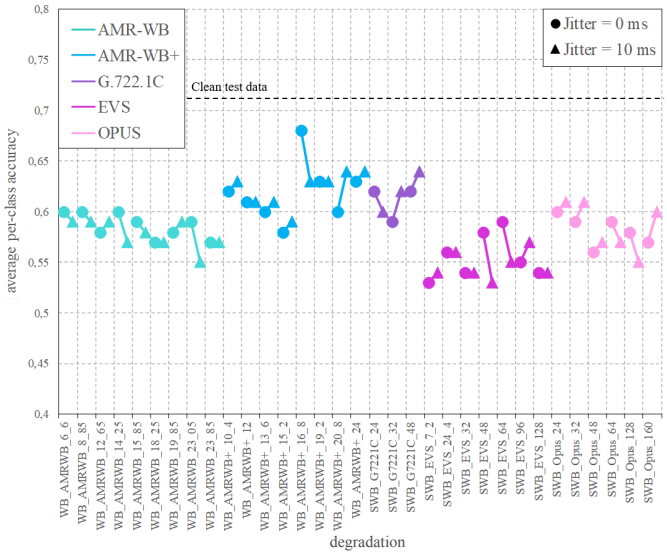

Fig. 7. Average per-class accuracy of SVC classifier with network-introduced impairments. PLR = 10%.

transmissions with the AMR-WB+ codec. These results refer to the case where no jitter was introduced, with the aim of isolating the effect of the packet loss.

There is a clear decrease in the performance of the RF classifier when the PLR is increased (Fig. 8). Despite this behavior, its performance is still better than that given by the SVC-rbf classifier, which presents more similar results with the growth of PLR (Fig. 9) as discussed previously. Regarding the coding bit-rate, there is no clearly observable trend in

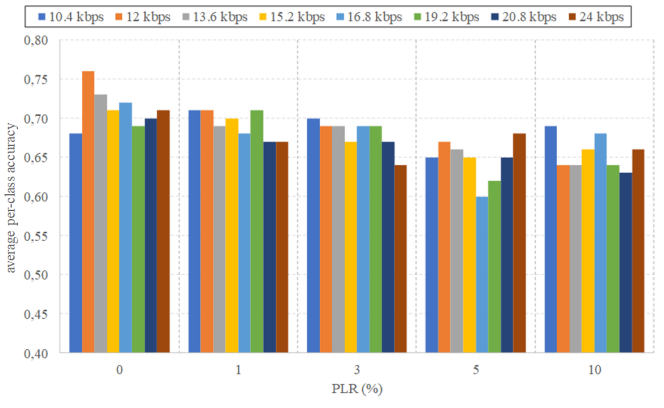

Fig. 8. Average per-class accuracy of both classifiers with the AMR-WB+ codec for different coding bit-rates (kbps) and PLR (%). RF classifier.

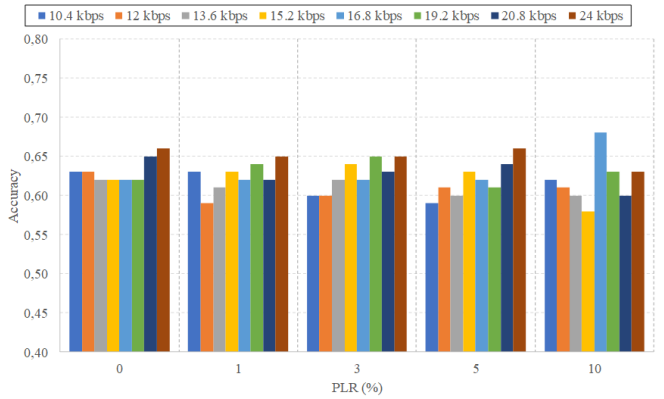

Fig. 9. Average per-class accuracy of both classifiers with the AMR-WB+ codec for different coding bit-rates (kbps) and PLR (%). SVC classifier.

any case. Hence, it cannot be concluded that increasing or decreasing coding bit-rate provides better performance in this classification task.

6 CONCLUSIONS

This work has focused on the automatic classification of speakers into two classes regarding their interpersonal characteristics manifested in the sound of their voices and manner of speaking: high *warmth–attractiveness* (WAAT) or low WAAT.

As speech data, we have considered a wide set of recorded voices which had been previously rated by a external panel of individuals. Relevant speech features have been extracted from speech segments and employed for training and testing binary classifiers.

The principal aim of this work is evaluating the impact that the impairments introduced by the encoding and communication processes has over the classification task. Thereby, an emulation test-bench has been deployed for

encoding and decoding the clean speech signals with a large set of speech codecs and, thereafter, transmitting it with controlled network-induced impairments, namely, jitter and packet loss.

With the mentioned binary classification task on the given data, we have found superior performance of two specific classifiers: one based on the Random Forest (RF) method and other based on Support Vector Machines with radial basis function kernel (SVC-rbf). Regarding speech coding without network transmission-related errors, it can be concluded that the use of narrowband codecs leads to very low accuracy of both classifiers, close to chance level. The AMR-WB (wideband codec), and the G.722.1C, EVS, and Opus (super-wideband codecs) offer more acceptable performance, while the AMR-WB+ (wideband codec) provides the best performance over all coded-decoded test speech signals. Only this codec offers performance close to that obtained with clean non-degraded speech. These findings are consistent for both classifiers tested.

The classifiers under study were only slightly affected by the appearance of jitter events, and more sensitive to the impact of packet loss, especially in the case of RF classification. However, under the considered conditions, the AMR-WB+ codec permitted accuracy levels over 0.70 with the RF classifier with moderate PLR (below 3%); in the case of unacceptable levels of PLR for voice-based communications, e.g. 10%, the accuracy reached by the RF classifier was over 0.65. The SVC-rbf classifier presented poorer performance but greater robustness against the effect of the impairments introduced by the network.

As future work, we plan to extend this study to other speaker's interpersonal characteristics, such as confidence, compliance, and maturity [Fernández Gallardo and Weiss 2018], as well as to improve the level of accuracy of the classification task by the refinement of the considered classifiers.

ACKNOWLEDGMENTS

This work has been supported by the German Research Foundation under Grant FE 1603/1-1 and by the European Union under the framework of the H2020 IoTCrawler project (contract 779852).

REFERENCES

A. Albahri, M. Lech, and E. Cheng. 2016. Effect of Speech Compression on the Automatic Recognition of Emotions. *International Journal of Signal Processing Systems* 4, 1 (2016), 55–61.

A. Aly and A. Tapus. 2013. A Model for Synthesizing a Combined Verbal and Nonverbal Behavior Based on Personality Traits in Human-Robot Interaction. 325–332.

A. Argal, S. Gupta, A. Modi, P. Pandey, S. Shim, and C. Choo. 2018. Intelligent Travel Chatbot for Predictive Recommendation in Echo Platform. In *Computing and Communication Workshop and Conference (CCWC)*.

E. Aronson, T. D. Wilson, and R. M. Akert. 2009. *Social Psychology* (7 ed.). Prentice Hall.

J. R. Bellegarda. 2014. *Natural Interaction with Robots, Knowbots and Smartphones*. Springer, New York, NY, Chapter Spoken Language Understanding for Natural Interaction: The Siri Experience, 3–14.

F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger. 2011. 'Would you Buy a Car From Me?' – On the Likability of Telephone Voices. In *Interspeech*. 1557–1560.

A. Burmania and C. Busso. 2017. A Stepwise Analysis of Aggregated Crowdsourced Labels Describing Multimodal Emotional Behaviors. In *Annual Conference of the International Speech Communication Association (Interspeech)*. 152–156.

A. S. Chowdhury and G. Riccardi. 2017. A Deep Learning Approach To Modeling Competitiveness In Spoken Conversations. In *ICASSP*. 5680–5684.

M. Eskénazi, G. A. Levow, H. Meng, G. Parent, and D. Suendermann. 2013. *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*. Wiley.

ETSI EG 202 396-2. 2006. *Speech Processing, Transmission and Quality Aspects (STQ); Speech quality performance in the presence of background noise; Part 2: Background noise transmission - Network simulation - Subjective test database and results*. European Telecommunications Standards Institute.

ETSI TR 103 138. 2016. *Speech and multimedia Transmission Quality (STQ); Speech samples and their use for QoS testing*. European Telecommunications Standards Institute.

F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* 7, 2

(2016), 190–202.

F. Eyben, F. Weninger, F. Gross, and B. Schuller. 2013. Recent Developments in OpenSMILE, the Munich Open-Source Multimedia Feature Extractor. In *ACM Multimedia (MM)*. 835–838.

L. Fernández Gallardo. 2016. *Human and Automatic Speaker Recognition over Telecommunication Channels*. Springer-Verlag, Singapore.

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

L. Fernández Gallardo, S. Möller, and J. G. Beerends. 2017. Predicting Automatic Speech Recognition Performance over Communication Channels from Instrumental Speech Quality and Intelligibility Scores. In *Interspeech*. 2939–2943.

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

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

ffmpeg. 2018. https://www.ffmpeg.org/

S. Frühholz, E. Marchi, and B. Schuller. 2016. The Effect of Narrow-band Transmission on Recognition of Paralinguistic Information from Human Vocalizations. *IEEE Access* 4 (2016), 6059–6072.

N. García, J. C. Vásquez-Correa, J. D. Arias-Londoño, J. F. Várgas-Bonilla, and J. R. Orozco-Arroyave. 2015. Automatic Emotion Recognition in Compressed Speech Using Acoustic and. Non-Linear Features. In *20th Symposium on Signal Processing, Images and Computer Vision (STSIVA)*.

J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost. 2017. Progressive Neural Networks for Transfer Learning in Emotion Recognition. In *Annual Conference of the International Speech Communication Association (Interspeech)*. 1098–1102.

Jan Holub, Michael Wallbaum, Noah Smith, and Hakob Avetisyan. 2018. Analysis of the Dependency of Call Duration on the Quality of VoIP Calls. *IEEE Wireless Communications Letters* 7, 4 (aug 2018), 638–641.

ITU-T Recommendation G.191. 2000. *Software Tools for Speech and Audio Coding Standardization*. International Telecommunication Union.

ITU-T Recommendation P.48. 1988. *Specification for an Intermediate Reference System*. International Telecommunication Union.

ITU-T Recommendation P.56. 1993. *Objective Measurement of Active Speech Level*. International Telecommunication Union, CH-Geneva.

I. Jacobs and W. Scholl. 2005. Interpersonale Adjektivliste (IAL). *Diagnostica – Zeitschrift für Psychologische Diagnostik und Differentielle Psychologie* 51, 3 (2005), 145–155.

D. Litman and K. Forbes-Riley. 2014. Evaluating a Spoken Dialogue System that Detects and Adapts to User Affective States. In *Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*. 181–185.

J. Masche and N.-T. Le. 2018. A Review of Technologies for Conversational Systems. In *Advances in Intelligent Systems and Computing*. Springer, 212–225.

G. Mohammadi and A. Vinciarelli. 2012. Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features. *IEEE Transactions on Affective Computing* 3, 3 (2012), 273–284.

S. Möller, A. Raake, N. Kitawaki, A. Takahashi, and M. Wältermann. 2006. Impairment Factor Framework for Wideband Speech Codecs. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 6 (2006), 1969–1976.

A. Ní Chasaide, I. Yanushevskaya, and C. Gobl. 2017. Voice-to-Affect Mapping: Inferences on Language Voice Baseline Settings. In *Annual Conference of the International Speech Communication Association (Interspeech)*. 1258–1262.

Q. Noirhomme, D. Lesenfants, F. Gomez, A. Soddu, J. Schrouff, G. Garraux, A. Luxen, C. Phillips, and S. Laureys. 2014. Biased Binomial Assessment of Cross-Validated Estimation of Classification Accuracies Illustrated in Diagnosis Predictions. *NeuroImage: Clinical* 4 (2014), 687–694.

S. Parthasarathy and C. Busso. 2017. Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning. In *Annual Conference of the International Speech Communication Association (Interspeech)*. 1103–1107.

J. Pohjalainen and P. Alku. 2014. Multi-Scale Modulation Filtering in. Automatic Detection of Emotions in Telephone Speech. In *ICASSP*. 980–984.

Ruhi Sarikaya. 2017. The Technology Behind Personal Digital Assistants: An overview of the system architecture and key components. *IEEE Signal Processing Magazine* 34, 1 (jan 2017), 67–81.

M. Schmitt, F. Ringeval, and B. Schuller. 2016. At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech. In *Annual Conference of the International Speech Communication Association (Interspeech)*. 495–499.

V. Silber-Varod, A. Lerner, and O. Jokisch. 2017. Automatic Speaker's Role Classification with a Bottom-up Acoustic Feature Selection. In *International Workshop on Grounding Language Understanding (GLU)*. 52–56.

B. Spillane, E. Gilmartin, C. Saam, K. Su, Cowan, B. R., S. Lawless, and V. Wade. 2017. Introducing ADELE: A Personalized Intelligent Companion. In *ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents*. 43–44.

Y. Teng and R. F. Kubichek. 2006. Speech Intelligibility Evaluation of Low Bit Rate Speech Codecs. In *12th Digital Signal Processing Workshop - 4th Signal Processing Education Workshop*. 251–256.

J. C. Vasquez-Correa, N. Garcia, J. R. Orozco-Arroyave, J. D. Arias-LondoÃśo, J. F. Vargas-Bonilla, and E. Nöth. 2015. Emotion Recognition from Speech under Environmental Noise Conditions using Wavelet Decomposition. In *IEEE International Carnahan Conference on Security Techology (ICCST)*. 247–252.

Xiaojie Wang and Caixia Yuan. 2016. Recent Advances on Human-Computer Dialogue. *CAAI Transactions on Intelligence Technology* 1, 4 (oct 2016), 303–312.

B. Weiss and F. Burkhardt. 2012. Is 'Not Bad' Good Enough? Aspects of Unknown Voices' Likability. In *Interspeech*. 510–513.

J. S. Wiggins, P. Trapnell, and N. Phillips. 1988. Psychometric and Geometric Characteristics of the Revised Interpersonal Adjective Scales (IAS-R). *Multivariate Behavioral Research* 23, 4 (1988), 517–530.

I. Zukerman and D. J. Litman. 2001. Natural Language Processing and User Modeling: Synergies and Limitations. *User Modeling and User-Adapted Interaction* 11, 1–2 (2001), 129–158.