Towards the Prediction of Human Speaker Identification Performance from Measured Speech Quality

Laura Fernández Gallardo[1,2], *Sebastian Möller*[2,1]

[1]Faculty of ESTeM, University of Canberra, Australia
[2]Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany
(laura.fernandez-gallardo|sebastian.moeller)@telekom.de

Abstract

Speech communication channels and their components (e.g. codecs) are generally designed for optimum perceived speech quality. However, transmission channels should also preserve principal speaker-specific characteristics that enable acceptable speaker identification performance by end listeners. This paper proposes a first step towards effective approaches for the prediction of the human speaker identification performance from instrumental quality measures. Correspondences between speech quality and speaker identification accuracy are shown by fitting linear curves to data points involving different channel transmissions. Narrowband, wideband, and super-wideband channels are considered, with other typically associated distortions. Our analyses show that *Coloration*, one of the perceptual quality dimensions, can be a better predictor of the human speaker identification performance than overall quality predictions in terms of Mean Opinion Scores. This suggests that the speaker-specific properties of the voice are mainly impaired by the distortion of frequency components in the transmission path.

Index Terms: human speaker identification, speech quality, instrumental measures, prediction model

1. Introduction

In the process of network planning and communication channel design, speech quality is an important factor to be considered in order to meet the users' needs and expectations. Human auditory tests are generally conducted to measure the quality perceived after speech transmission through the channel under development. Instrumental (objective) models (e.g. PESQ [1], POLQA [2] and E-model [3]) can partially replace such tests in order to get quick predictions of the auditory speech quality.

Speech technology engineers may not only aim to achieve good speech quality, but also to develop communication systems enabling acceptable human and automatic speech recognition and detection of the speaker identity and other paralinguistic information. Hence, it would also be interesting to consider other aspects as criteria for communication channel design, such as satisfactory speech recognition, speaker recognition, and/or speaker characterization (e.g. detection of age, emotion, personality, etc.), performance by humans or by automatic recognition algorithms. So far, these criteria are not systematically taken into account in the communication channel design process.

This paper investigates the relationships between speech quality and human speaker identification accuracies, when speech is transmitted over different channels. Correspondences are found by fitting models that permit the estimation of human speaker recognition performances from instrumental quality measures calculated from degraded speech signals. Different bandwidths, codecs, and electro-acoustic user interfaces are considered in this work since they are the main speech communications impairments. The relationships found empirically through our analyses may be useful for network planning purposes when the effects of the transmission channel need to be evaluated, but when conducting human and automatic speaker recognition tests is too costly in terms of time and resources.

Study Group 12 of the International Telecommunication Union (ITU-T) is currently studying the definition of a universal scale which would permit the comparison amongst test conditions and algorithms predicting different speech quality aspects. Our work is contributing to this aim by predicting human speaker identification performance [4, 5]. Only models for predicting overall quality [1, 2, 3], perceptual quality dimensions [6], intelligibility ([7, 8], although not yet satisfactory), and automatic speech recognition [9, 10] have been developed so far for telephone channels. However, to the best of our knowledge, no model for the prediction of speaker identification performance has been proposed before. The present research could also lead to the development of efficient algorithms for the prediction of automatic speaker recognition scores.

2. Communication Channels

Transmission channels of different bandwidths are considered in this work. The traditional narrowband (NB, 300–3,400 Hz), still predominant in the PSTN (Public Switched Telephone Network), offers poor perceived voice quality due to the limited transmission frequency range [11]. Differently, wideband (WB, 50–7,000 Hz) communications, generally found in Voice over IP (VoIP) applications, offer an extended signal bandwidth which is responsible for improved speech quality — an improvement of 29% has been found compared to NB [12]. Also the currently emerging super-wideband channels (SWB, 50–14,000 Hz) are studied in this work, which are preferred for video conferencing. The work in [13] reported that SWB offers 39% increased quality in comparison to WB and 79% in comparison to NB. The enhanced bandwidths also provide a significantly better recognition of the talker by listeners and by automatic systems compared to their performance in NB [14].

Besides the bandwidth filter, other artefacts of the transmission channel such as the employed coding algorithm, the bit rate, the electro-acoustic user interfaces at both ends of the communication, and packet loss introduce main impairments into the signal and affect the perceived quality to different extents.

Channels offering better quality enable more accurate speaker recognition in most cases. A comparative table shown in [11] permits the classification of transmission channels into eight service quality classes, which may be the basis for transmission network planning.

3. Human Speaker Identification Tests

The human speaker identification performance was tested by performing a series of auditory tests. The voice of 16 speakers (8m, 8f) was recorded in clean conditions (in an acoustically isolated room, 48 kHz sampling frequency, and employing a high-quality microphone and sound card). Segments of different lengths (words, and sequences of three words) were extracted from the recordings and degraded according to different channel conditions examined in two listening tests. The tests consisted in the identification of familiar speakers—all listeners were acquainted with the speakers—in a closed-set setup. The stimuli were randomized and presented in different order to each listener through high-quality headphones. The task was to select one of the 16 possible speakers after listening to each stimulus. The listeners and the speakers speak German as mother tongue and have an age range of 24–47 years.

In the first listening test (LT1) conducted, the audio material were the German words *auch*, *immer* and *können*, meaning *also*, *always*, and *be able to*, respectively. These words were processed with three NB codecs (G.711 at 64 kbit/s, GSM-EFR at 12.2 kbit/s, and AMR-NB at 4.75 kbit/s) and two WB codecs (G.722 at 64 kbit/s and AMR-WB at 23.05 kbit/s), creating five versions of the audio segments. The corresponding channel filter (ITU-T Rec. G.712 for NB and ITU-T P.341 for WB) and codec were applied via software simulation.

The segment in German *Könnten Sie mir*, meaning *Could you () me* (start of sentence) was degraded with different conditions for the second listening test (LT2). The signals were transmitted through different terminals in sending direction (an IP phone with handset, a mobile phone, a hands-free phone and a headset) employing a head-and-torso simulator. The examined codecs were G.711 at 64 kbit/s and AMR-NB at 12.2 kbit/s in NB, G.722 at 64 kbit/s and AMR-WB at 12.65 kbit/s in WB, and G.722.1C at 32 and at 48 kbit/s in SWB.

All the test conditions are listed in the first column of Table 2. 26 listeners participated in LT1 and 20 in LT2. More details about the test procedures and results are reported in [15, 16].

4. Instrumental Quality Measurements

The speech quality of telephone transmissions with different settings can be estimated employing a variety of instrumental techniques. This work considers the signal-based models POLQA [2] and DIAL [6]. Mean opinion scores (MOS) were estimated on a joint scale in the range [1–4.5]. These instrumental measures were applied to obtain the estimated quality of each of the communication channels of Section 3. Longer segments of the same speakers and degradations as those of the described listening tests were employed for the speech quality measurements.

The SWB mode for POLQA was employed with the objective of a comparison of the NB, WB, and SWB conditions. The DIAL model provides, in addition to MOS, estimations of the four perceptual quality dimensions *Coloration*, *Discontinuity*, *Noisiness* and (sub-optimal) *Loudness*; these dimension estimates can serve as a diagnostic cause analysis of the estimated degradations [6].

The formal term of the *Coloration* perceptual dimension is ”Directness/Frequency Content” (DFC). It can be regarded as the distortion of the speech frequency components caused by the mouth-to-ear transmission path. This perceptual dimension is directly influenced by electro-acoustic user interfaces, network bandwidth, and acoustical properties of the speaker's and listener's environment, such as room reflections [6]. The investigation in [17] describes an estimate of the DFC in terms of two parameters of a transmission system's gain function $G'(\Omega)$: the bandwidth and the center of gravity θ_G of $G'(\Omega)$.

Other instrumental quality measures such as PESQ [1] and E-model [3] are not considered in the present study as they are not applicable to channels involving electro-acoustic interfaces. In addition, POLQA and DIAL are more advanced models than PESQ and are adapted to a wider variety of network scenarios [6].

The instrumental quality estimation with POLQA and DIAL involved the use of a reference file without any degradation. Requirements regarding pre-processing of reference and of degraded files can be found for POLQA in ITU-T Rec. P.863 (Sections 8.1 and 8.3). For DIAL, the same pre-processing as for POLQA was applied. According to these requirements, the reference files were sampled at 48 kHz, band-pass-filtered to 50–14,000 Hz with the filter of ITU-T Rec. P.341, and level-equalized 26 dB below the overload of the digital system (-26 dBov) by applying the ITU-T Rec. P.56. The pre-processing applied to the degraded files was: band-pass-filtering to 50–14,000 Hz, resampling to 48 kHz, and level-equalization to -26 dBov as indicated before.

5. Predicting Human Speaker Identification Performance

Various models are fitted to observations of measured speech quality and human speaker identification accuracies and the best fit is then determined. All curves in this work are fitted by solving linear least squares problems using the QR factorization. For each transmission condition, the average of the listeners' accuracies was calculated and the corresponding speech quality computed. POLQA MOS and DIAL *Coloration* are employed as estimators of the listeners' performance.

5.1. POLQA MOS as Estimator

A first-degree polynomial curve $l1(x)$ was fitted to the pairs *speech quality–speaker identification accuracy from words (of LT1)*. This fit is shown in Figure 1. The 95% confidence intervals of the fitted curve are plotted with discontinuous lines—as in all curve plots of this paper—according to the range within which the curve coefficients have been estimated. The speech quality was represented by the MOS given by POLQA. A linear curve was chosen in order to avoid overfitting. It is shown in [14] that quadratic curves are less useful for prediction in this case, due to the low number of data points considered.

The goodness of the fit is evaluated from the R-squared value (R^2) and from the Root Mean Squared Error (RMSE). The RMSE was calculated as the square root of the Mean Square Error (MSE), which is defined as the sum of the squares of the residuals divided by the degrees of freedom.

Next, it is shown that the curve $l1(x)$ can be employed to predict the human speaker identification accuracies of a different scenario (LT2), when other communication channels are employed for voice transmission. The speaker identification rates corresponding to these speech files can be predicted well with the computed $l1(x)$ when its intercept parameter $b1$ is allowed to vary—the new curve with a different $b1$ is referred to

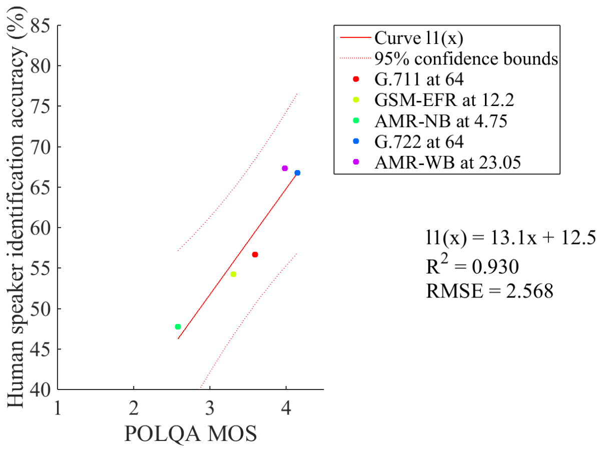

Figure 1: *Linear polynomial curve $l1(x)$ fitted to the pairs POLQA MOS—human speaker identification from words.*

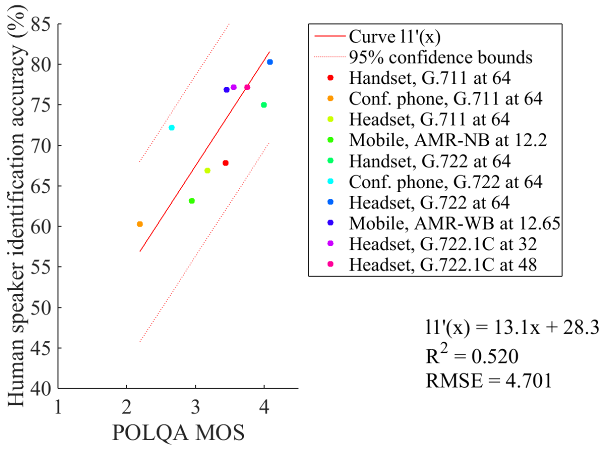

Figure 2: *Curve $l1'(x)$ to predict human speaker identification rates from segments transmitted through various user interfaces from POLQA MOS values.*

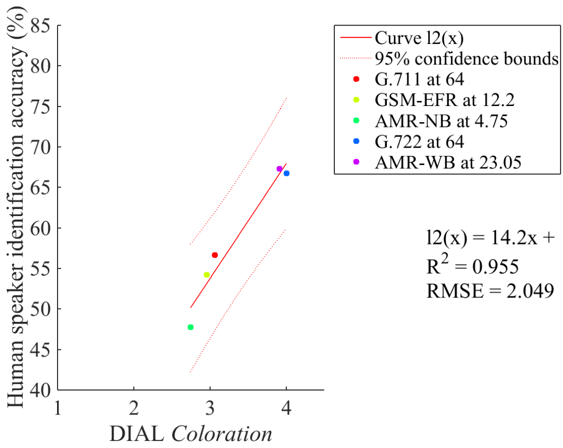

Figure 3: *Linear polynomial curve $l2(x)$ fitted to the pairs DIAL Coloration—human speaker identification from words.*

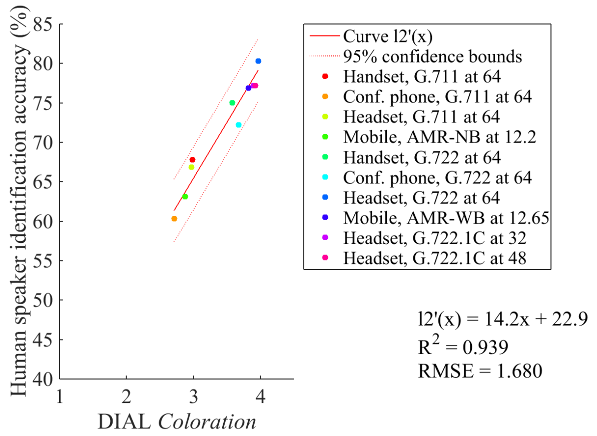

Figure 4: *Curve $l2'(x)$ to predict human speaker identification rates from segments transmitted through various user interfaces from DIAL Coloration values.*

as $l1'(x)$, and the varied parameter to as $b1'$. The fit of $l1'(x)$ is shown in Figure 2.

The variations in the intercept parameter account for the alteration of the range within which the accuracies of the new listening test are obtained, shifting the curve up or down but maintaining constant the curve slope. The test accuracies heavily depend on the length of the stimuli heard and, presumably to a lesser extent, on the phonological content and on the familiarity and distinctiveness of the voices of the test. The number of speakers in the identification task is probably an important factor too, although this number was the same for the two tests (it can be assumed that the number of talkers to be identified in real phone calls remains unaltered across different conditions and communication distortions).

In this case, the factor causing higher identification rates in LT2 compared to LT1 was assumed to be the length of the stimuli. Fixing the slope parameter $a1$ to its value $a1$ = 13.1, the value of the new $b1'$ parameter which best fitted the accuracies of the second test was $b1'$ = 28.3, larger than $b1$ of the initial model fit. The new model accounted for only 52% of the data variance (R^2 = 0.520). This weak fit is partly due to the high variability of the quality–speaker identification pairs across the different user interfaces, as can be observed in the plot of Figure 2. For instance, the expected identification accuracies for speech of low quality such as that resulting from the transmission through the hands-free terminal in NB (POLQA MOS = 2.19) are within the interval [45.71%; 68.02%], whereas the accuracies for speech offering high quality as offered by the headset in WB (POLQA MOS = 4.08) are expected within [70.40%; 92.71%], with 95% confidence level.

5.2. DIAL *Coloration* as Estimator

Interestingly, a model built from the *Coloration* perceptual dimension values given by DIAL has been found to fit the data better than that built from the POLQA MOS. A linear polynomial curve fitted to the *Coloration–speaker identification results of LT1*, termed $l2(x)$, is shown in Figure 3.

As done previously, the slope parameter was fixed to the value found ($a2$ = 14.2) and the intercept parameter $b2$ allowed to vary to align the curve to the range of identification scores of the second listening test. The new computed curve is referred to as $l2'(x)$, shown in Figure 4, and the value of its independent parameter is $b2'$ = 22.9. This is a better fit than that obtained with $l1'(x)$, which considered POLQA MOS as estimator (shown in Figure 2). It can be seen in Figure 4 that the 95% confidence bounds of $l2'(x)$ are found around [-4.0%; +4.0%] the estimated value. This range is about three times narrower than the range [-11.2%; +11.2%] of $l1'(x)$, observed in Figure 2.

A comparison between the different fitted curves and their figures of merit (R^2 and RMSE) is given in Table 1, in which also the models obtained with DIAL MOS as predictor (l3(x) and l3'(x)) are shown. These curves were obtained following a similar process as for $l1(x)$, $l1'(x)$, and $l2(x)$, $l2'(x)$. DIAL MOS as predictor provides a weaker fit to the data of the second listening test compared to that obtained from DIAL *Coloration*.

The other perceptual dimensions (*Discontinuity*, *Noisiness*, and *Loudness*) were not found to be valid estimators of human speaker recognition scores, since very weak or no curve fits could be established. Differently, it seems that the effects of the user interfaces are well reflected by DIAL *Coloration*. Because the human speaker recognition performance is also influenced by the transmission through different terminals [16], it results that *Coloration* is a good estimator of the human speaker identification accuracies. The estimated overall MOS quality takes also into account the other three perceptual dimensions, less affected by transmitting devices, and hence offers less accurate prediction models in comparison to DIAL *Coloration*. Therefore, the listeners' accuracy can be better approximated from *Coloration* values than from other perceptual dimensions or the MOS quality when user interfaces are involved in the transmission.

Table 1: *Linear polynomial curves and figures of merit.* $l1(x)$ *and* $l2(x)$ *were fitted to LT1 data (NB, WB, no user interface), and* $l1'(x)$ *and* $l2'(x)$ *to LT2 data (NB, WB, SWB, 4 different user interfaces). The slope coefficient obtained from the fit to the LT1 data were fixed for the fit to the LT2 data.*

Predictor	Curve	R^2	RMSE
POLQA MOS	$l1(x) = 13.1x + 12.5$	0.930	2.568
	$l1'(x) = 13.1x + 28.3$	0.520	4.701
DIAL *Coloration*	$l2(x) = 14.2x + 11.2$	0.955	2.049
	$l2'(x) = 14.2x + 22.9$	0.939	1.680
DIAL MOS	$l3(x) = 21.8x - 14.0$	0.988	1.046
	$l3'(x) = 21.8x - 3.4$	0.556	4.521

6. Using the Best Fits for Predictions

It has been assessed that the models built from DIAL *Coloration* values are the best predictors of human speaker identification scores under transmissions involving different sending devices. For some transmission and stimuli length conditions for which no listening test was conducted [15, 16], $l2(x)$ was employed to predict human speaker identification accuracies from words, and $l2'(x)$ to predict the accuracies from starts of sentences. The new computed values are shown in Table 2, along with real accuracies resulting from the auditory tests for a different stimulus length. The new accuracies are predicted within a range of [-7.7%; +7.7%] on average in the case of word stimuli and within a range of [-4.0%; +4.0%] on average for starts of sentences, with 95% confidence level. More accurate models with narrower prediction intervals could have possibly been found with more data points in LT1, i.e. employing more transmission conditions. This leaves room for future improvement.

Table 2: *True accuracies obtained from the listening tests and values (shaded) predicted with the model* $l2(x)$ *(for words, LT1) or* $l2'(x)$ *(for starts of sentences, LT2). The DIAL Coloration (DIAL C) measures are the predictors.*

Transmission Channel	DIAL *C*	Acc.(%) LT1	Acc.(%) LT2
No device, G.711 at 64	3.07	56.65	66.45
No device, GSM-EFR at 12.2	2.96	54.25	64.89
No device, AMR-NB at 4.75	2.75	47.76	61.92
Handset, G.711 at 64	2.99	53.60	67.81
Conf. phone, G.711 at 64	2.71	49.63	60.31
Headset, G.711 at 64	2.97	53.32	66.88
Mobile, AMR-NB at 12.2	2.88	52.04	63.13
No device, G.722 at 64	4.01	66.75	79.77
No device, AMR-WB at 23.05	3.91	67.31	78.36
Handset, G.722 at 64	3.58	61.96	75.00
Conf. phone, G.722 at 64	3.67	63.24	72.19
Headset, G.722 at 64	3.97	67.49	80.31
Mobile, AMR-WB at 12.65	3.82	65.36	76.88
Headset, G.722.1C at 32	3.89	66.36	77.19
Headset, G.722.1C at 48	3.93	66.92	77.19

7. Conclusions

To summarize, the linear polynomial curves $l2(x)$ and $l2'(x)$ have been fitted to observations of human speaker identification accuracies from speech with different degradations and the corresponding DIAL *Coloration* measures. The goodness of these fits was found to be the best compared to fits that employ other quality measures. Relatively lower R^2 and higher RMSE values have been found assessing the fits of $l1(x)$ and $l1'(x)$, which were estimated in the same way as $l2(x)$ and $l2'(x)$, respectively, yet requiring POLQA MOS values as input. No strong correspondences were found between DIAL MOS (or with DIAL perceptual dimensions other than *Coloration*) and speaker identification rates either. The models $l2(x)$ and $l2'(x)$ facilitate the prediction of the human speaker identification performance and may be useful for transmission planners to select from different possible network configurations.

The fact that the DIAL predictions of *Coloration* can be mapped to human speaker identification performance more satisfactorily than MOS is an interesting outcome. This indicates that the speaker-specific properties of the voice spectrum are impaired by *Coloration* to a greater extent than by the other quality dimensions separately (such as *Discontinuity*, *Noisiness*, and *Loudness*), and also than by the integral quality (MOS).

In future work, additional listening tests considering different stimulus contents and a wider range of telephone degradations can be conducted in order to obtain a more descriptive set of human speaker recognition or speaker characterization performances. The ultimate goal of this research is the creation of a useful tool to facilitate the design of communication channels taking into account different aspects of the speech signal other than its overall quality. This would result in a significant reduction of the costs of performance evaluations via listening tests.

8. References

[1] ITU-T Recommendation P.862.2, *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*, International Telecommunication Union, CH-Geneva, 2007.

[2] ITU-T Recommendation P.863, *Perceptual Objective Listening Quality Assessment*, International Telecommunication Union, CH-Geneva, 2011.

[3] ITU-T Recommendation G.107.1, *Wideband E-Model*, International Telecommunication Union, CH-Geneva, 2011.

[4] ITU-T Temporary Document TD 65 Rev. 1, *Status Report for Question 8/12*, Source: Rapporteur Q.8/12 (Author: S. Möller), ITU-T SG12 Meeting, CH-Geneva, 19–28 March, 2013.

[5] ITU-T Contribution COM 12-41, *Comparability of Quality Indices on the MOS and the R Scale*, Source: Deutsche Telekom AG (Author: S. Möller), ITU-T SG12 Meeting, CH-Geneva, 19–28 March, 2013.

[6] N. Côté, *Integral and Diagnostic Intrusive Prediction of Speech Quality*. Berlin, Germany: Springer, 2011.

[7] J. G. Beerends, S. van Wijngaarden, and R. van Buuren, "Extension of ITU-T Recommendation P.862 PESQ towards Measuring Speech Intelligibility with Vocoders," in *New Directions for Improving Audio Effectiveness. Meeting RTO-MP-HFM-123*, 2005, pp. 10–1–10–6.

[8] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, "An Evaluation of Objective Quality Measures for Speech Intelligibility Prediction," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2009, pp. 1947–1950.

[9] H. Sun, L. Shue, and J. Chen, "Investigations into the Relationship Between Measurable Speech Quality and Speech Recognition Rate for Telephony Speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2004, pp. 865–868.

[10] A. C. M. Rietveld and A. P. A. Broeders, "Testing the Fairness of Voice Identity Parades: the Similarity Criterion," in *International Congress of Phonetic Sciences (ICPhS)*, 1991, pp. 46–49.

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

[12] S. Möller, A. Raake, N. Kitawaki, A. Takahashi, and M. Wältermann, "Impairment Factor Framework for Wideband Speech Codecs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1969–1976, 2006.

[13] M. Wältermann, I. Tucker, A. Raake, and S. Möller, "Extension of the E-Model Towards Super-Wideband Speech Transmission," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 4654–4657.

[14] L. Fernández Gallardo, "Human and Automatic Speaker Recognition over Telecommunication Channels," Ph.D. dissertation, University of Canberra, Australia, submitted (2014).

[15] L. Fernández Gallardo, S. Möller, and M. Wagner, "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*, 2012, pp. 219–222.

[16] L. Fernández Gallardo, S. Möller, and M. Wagner, "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)*, 2013, pp. 7775–7779.

[17] K. Scholz, M. Wältermann, L. Huo, A. Raake, S. Möller, and U. Heute, "Estimation of the Quality Dimension 'Directness/Frequency Content' for the Instrumental Assessment of Speech Quality," in *International Conference on Spoken Language Processing (ICSLP)*, 2006, pp. 1523–1526.