Human Speech Intelligibility Measurements over VoIP Channels

Laura Fernández Gallardo

Quality and Usability Lab, Telekom Innovation Labs, TU Berlin, Deutschland, Email: laura.fernandezgallardo@tu-berlin.de

Abstract

VoIP (voice over Internet Protocol) transmissions are able to deliver wideband (50-7,000 Hz) and super-wideband (50-14,000 Hz) voice, which brings manifold advantages over conventional narrowband channels (300-3,400 Hz). It has been found that, while perceived speech quality, speaker recognition, and automatic speech recognition performance are improved, also human speech intelligibility can benefit from the extended bandwidths. However, it still remains unclear whether the human intelligibility can be enhanced with super-wideband and full-band quality with respect to wideband, and whether it is significantly affected by coded-decoded speech. This paper presents an auditory test conducted with 30 participants where their intelligibility was assessed over 27 channels of different bandwidths, codecs, and bit rates. This test was based on a closed set of vowel-consonant-vowel logatomes with eight alternatives. Furthermore, it is shown that the subjective intelligibility scores can be well predicted by POLQA mean opinion scores and POLQA-based intelligibility objective measures.

Introduction

In recent years, we have witnessed substantial deployment of digital communication channels. Conventional narrowband (NB) telephony took advantage of the lower frequencies of the voice spectrum, promoting the standard 300-3,400 Hz bandwidth of the public switched telephone network (PSTN). A number of voice over internet protocol (VoIP) channels were more recently developed to deliver narrowband and wideband (WB) speech, the latter being the result of expanding the NB bandwidth to 50-7,000 Hz, offering a more natural voice. Also, to satisfy the demand of even higher quality for other signals, such as music, efforts have been made towards extending WB codecs to super-wideband (SWB) and full-band (FB) codecs, operating in the 50-14,000 Hz and 20-20,000 Hz frequency ranges, respectively [1, 2].

It has been shown that widening the telephony spectrum from NB to WB results in about 30% improvement in perceptual speech quality [3]. It was later found in [4] that SWB offers 39% increased quality in comparison to WB and 79% in comparison to NB. The extended bandwidth also allows for improved speaker recognizability. Results from several listening tests suggest that known speakers can be easier identified in WB compared to NB [5], and advantages of SWB channels over NB and WB have been found for automatic speaker recognition [6](Section 5.2.1). Automatic speech recognition also benefits from the enhanced communication bandwidths [7, 8].

The mentioned advantages and the fact that many high-frequency sounds are critical for human speech intelligibility [9] can imply an improved intelligibility performance when listening to voices in the extended bandwidths compared to NB [10]. However, only one study exists, to the best of our knowledge, that examined the difference in phoneme intelligibility performance over different bandwidths [11]. That study detected a superior performance in WB (G.722 codec) compared to NB (AMR-NB codec). Other investigations have only considered the effect of NB degradations [12, 13] or of codecs operating on speech sampled at 16 kHz [14]. It still remains to be shown whether the superiority of WB holds for a wider variety of codecs and bitrates, and whether the switch from WB to SWB transmissions brings a further increase in intelligibility performance.

The present contribution reports a listening test to measure human speech intelligibility, where the speech stimuli were degraded through 23 channel distortions involving different bandwidth filtering, coding schemes, and bitrates. The original speech signals and three down-sampled versions were also considered as experiment conditions. Our main objective was to assess the differences in intelligibility performance over NB, WB, and SWB, while possibly identifying particularities in the performance delivered by some codecs or bitrates. A listening test employing a closed set of nonsense vowel-consonant-vowel (VCV) logatomes was considered as suitable for this investigation. Since consonants are crucial for intelligible speech, we opted for employing monosyllabic stimuli varying consonants (a middle consonant) and maintaining unchanged the enclosing vowel sounds. Other intelligibility studies also employing non-sense combinations of vowels and consonants examine the effects of noise [15, 16] or of bandwidth-filtering [9].

The choice of a logatome-based intelligibility test enabled us to compare the subjective results to objective predictions given by the POLQA intelligibility model (V1490intellV2). This model is a further development of PESQ intelligibility [17] using the latest developments in the objective assessment of speech quality [18, 19]. Our intelligibility results were also compared to objective transmitted speech quality predictions made by the POLQA standard V2.4.1 (objective MOS).

Audio material

Eight different VCV logatomes were chosen for the intelligibility test, varying the middle consonant: "ama", "aba", "afa", "ana", "apa", "asa", "awa", and "ascha". The choice of these logatomes was based on the high phoneme confusions previously found in [11].

The logatomes were extracted from words in purposely created sentences recorded by 4 German speakers (2m, 2f, age range 25–36 years). The recordings were made with sampling frequency of 48 kHz in clean conditions. The test stimuli are thus excerpts of natural speech, which can presumably be less carefully articulated than words or logatomes spoken in isolation as in the OLLO logatome speech database [16, 11], but on the other hand reflect a realistic pronunciation.

The 32 excerpts (4 speakers x 8 logatomes) were transmitted through 23 channel conditions via software simulation. These include uncoded and coded-decoded speech of three telephone bandwidth (NB, WB, and SWB) at different bitrates. Besides, conditions with no distortion applied, i.e. direct speech sampled as 8, 16, 32, or 48 kHz were examined. The condition names can be seen in Figure 1, which displays the intelligibility test results.

The channel transmissions involved downsampling the speech signal with an anti-aliasing filter to 8, 16, or 32 kHz for the NB, WB, or SWB conditions, respectively. The speech was then level-equalized 26 dB below the overload of the digital system (-26 dBov), a characteristic level of telephone channels, using the voltmeter algorithm of ITU-T Rec. P.56. For the channel degradations, a bandwidth filter was applied, complying with ITU-T Rec. G.712 for NB, ITU-T Rec. P.341 for WB, and 14KBP for SWB. Finally, codecs were applied at different bitrates for the conditions involving a codec, and the speech was again level-equalized to -26 dBov.

Subjective intelligibility ratings

Intelligibility test setup

The complete set of stimuli presented in the intelligibility test consisted of 4 speakers x 8 VCV logatomes x 27 conditions = 864 segments. The task for the test participants was to choose among the eight logatome alternatives after listening to each stimulus. There was no possibility to listen to the stimuli more than once. Short breaks were included every 15 minutes approximately to avoid listeners' loss of focus. A brief familiarization phase was conducted before the actual test started. In the familiarization, the listeners clicked on each logatome button to hear a sample as many times as they wished.

The test was performed by 30 listeners (15m, 15f), mean age 25 years (range 18–38 years) and with German as mother tongue. The complete test session had a duration of about one hour. It was performed in a 54m^2 acoustically treated listening room (room Pinta in the Telefunken building of TU Berlin) using a laptop and Shure SRH240 headphones (diotic listening, frequency range 20–20,000 Hz). Listeners were not allowed to control the speech loudness level.

Intelligibility test results

The performance of the group of listeners was computed as the percentage of correct answers calculated over all

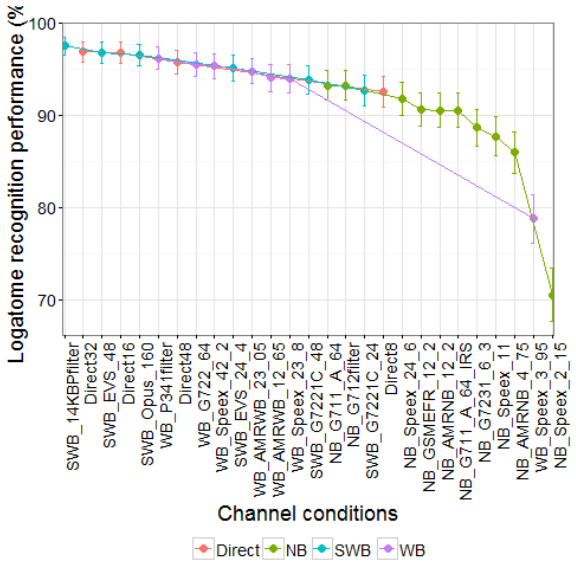

Figure 1: Human intelligibility accuracy across channel conditions.

speakers and logatomes for each condition. Figure 1 displays the obtained accuracies sorted from high to low.

It can be observed that better intelligibility performance is obtained with speech of greater bandwidth and with codecs operating at a higher bitrate. Consistently for each bandwidth, the condition which involves a bandwidth filter and no codec offers better intelligibility accuracy than the rest of channels of the same bandwidth.

The non-parametric Kruskal-Wallis test was applied in order to detect statistically significant differences in performance ($p < .05$) comparing the channel conditions. With respect to the benefits of the switch from NB to WB, the three WB conditions WB-Speex at 42.2 kbit/s, G.722 at 64 kbit/s, and P.341filter permit a significantly better performance compared to the tested NB codecs at a bit rate of 11 kbit/s or lower. The rest of WB conditions (except for WB-Speex at 3.95 kbit/s) offer significantly better performance than that of AMR-NB at 4.75 kbit/s and of Speex at 2.15 kbit/s. In view of the test results, the significant advantages of the extended bandwidth are only manifested with WB codecs operating at a sufficient bitrate with respect to NB codecs at low bitrates. The performance of the Speex codec is markedly worse than other codecs at similar bitrates in NB and in WB. Other disadvantages of the Speex codec have previously been shown in [20, 8].

Among the conditions tested, the SWB performance is statistically similar to that in WB, except for the WB-Speex at 3.95 kbit/s. This suggests that the relevant frequencies contributing to human intelligibility are already included in the WB bandwidth. Comparing SWB to NB, the worst performing SWB conditions, G.722.1C at 24 and at 48 kbit/s, are only significantly better than NB-Speex at 2.15 kbit/s. The best performing condition, 14KBPfilter, offers significantly higher accuracy than all NB conditions except for G.711 at 64 kbit/s and G.712filter.

High rates of logatome confusions have been found between "aba"-"awa" (in both directions, "w" being better recognized that "b"), and "afa"-"asa" (when "afa" was presented). A remarkable reduction of confusions could be identified with the switch from NB to WB, especially for "awa"-"aba" (when "awa" was presented) and for "afa"-"asa" (when "afa" was presented), as also found in [11]. This decrease of confusions between "afa"-"asa" was expected, as the phones "f" and "s" present similar spectral characteristics in the lower frequencies but spectral peaks at different distinctive locations over 6 kHz [21, 22]. High confusion rates between "aba"-"awa" still prevail for SWB channels and "Direct" conditions with $fs >= 16\ kHz$. The voiced bilabial stop "b" and the voiced labiodental fricative "w" appear to have spectral similarities [21] with decreasing energy until approximately 3 kHz.

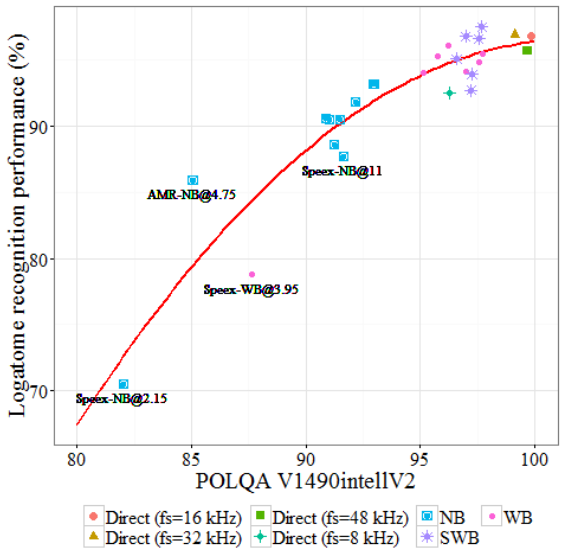

Figure 2: Subjective intelligibility scores vs. POLQA-intelligibility model estimations and second-order fit. $q_{POLQAINTELL}(x) = 92.1 + 27.2x - 7.6x^2$, $R^2 = 0.870$, $RMSE = 2.10$.

Subjective and objective intelligibility

The subjective intelligibility scores obtained in the test are compared to the objective scores provided by the POLQA intelligibility model (V1490intellV2) and to the MOS provided by the POLQA standard (V2.4.1). Both models operated with an input speech file containing the eight logatomes concatenated, uttered by one speaker. The models were applied for each channel condition and the resulting scores were then averaged across the four speakers.

Our results reveal that a second-order curve can be fit to the pairs subjective vs. objective measures with a remarkably high R^2. The fit is slightly better when the subjective intelligibility predictions are made by the POLQA intelligibility model compared to when those are made by the POLQA MOS. This possibility of predicting subjective test results might be useful for network engineers in the communication channel design process, when subjective testing costs are prohibitively high.

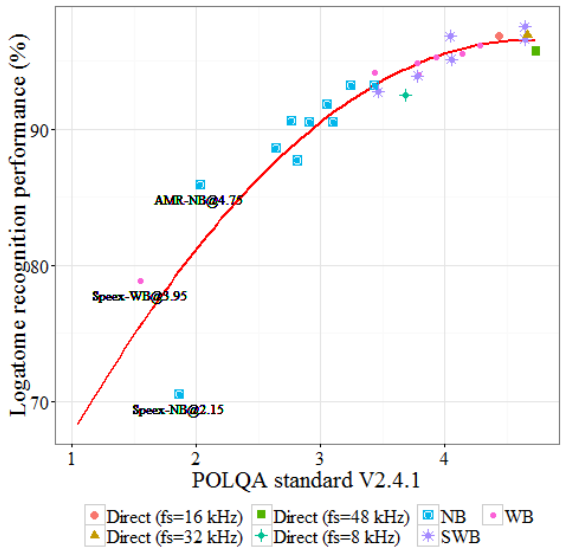

Figure 3: Subjective intelligibility scores vs. POLQA MOS estimations and second-order fit. $q_{POLQAMOS}(x) = 92.1 + 26.5x - 9.3x^2$, $R^2 = 0.858$, $RMSE = 2.20$.

The curves and data points are plotted in Figures 2 and 3, for POLQA intelligibility and for POLQA MOS, respectively. The displayed texts indicate data points detected as potential outliers, with high leverage or large absolute residual value. According to Figure 2, the AMR-NB codec at 4.75 kbit/s offers a greater subjective intelligibility score as predicted by the model, with higher residual value than other data points, in contrast to the outliers corresponding to the Speex codec. Figure 3 exhibits that the relation between human intelligibility and quality generally holds for the channel conditions tested.

Conclusions

A closed-response intelligibility test employing eight VCV logatomes was conducted with 30 listeners. The speech stimuli were degraded by 23 channel conditions involving NB, WB, and SWB, and included undistorted and downsampled speech. The gain in intelligibility accuracy is more salient for the transition from NB to WB than from WB to SWB, which indicates that the frequency components critical for consonant intelligibility are found in the bandwidth (50–7,000 Hz). Significant differences ($p < .05$) have only been found comparing WB conditions at a high bitrate to NB conditions at a low bitrate.

The quadratic correspondences between subjective and objective intelligibility scores have been shown to be high, with $R^2 = 0.870$; $RMSE = 2.10$ and $R^2 = 0.858$; $RMSE = 2.20$ when the objective scores were predicted by POLQA intelligibility (V1490intellV2) and by POLQA MOS (V2.4.1), respectively. This result highlights the goodness of these objective measures to estimate subjective intelligibility scores.

Acknowledgements

The author would like to thank Mr. John Beerends for his valuable support and for kindly providing the objective intelligibility scores.

References

[1] M. Tammi, L. Laaksonen, A. Rämö, and H. Toukomaa, "Scalable Superwideband Extension for Wideband Coding," in *ICASSP*, 2009, pp. 161–164.

[2] J.-M. Valin, T. B. Terriberry, and G. Maxwell, "A Full Bandwidth Audio Codec with Low Complexity and Very Low Delay," in *European Signal Processing Conference (EUSIPCO)*, 2009, pp. 1254–1258.

[3] 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.

[4] M. Wältermann, I. Tucker, A. Raake, and S. Möller, "Extension of the E-Model Towards Super-Wideband Speech Transmission," in *ICASSP*, 2010, pp. 4654–4657.

[5] L. Fernández Gallardo, S. Möller, and M. Wagner, "Human Speaker Identification of Known Voices Transmitted Through Different User Interfaces and Transmission Channels," in *ICASSP*, 2013, pp. 7775–7779.

[6] L. Fernández Gallardo, *Human and Automatic Speaker Recognition over Telecommunication Channels*, ser. T-Labs Series in Telecommunication Services. Singapore: Springer-Verlag, 2016.

[7] A. V. Ramana, L. Parayitam, and M. S. Pala, "Investigation of Automatic Speech Recognition Performance and Mean Opinion Scores for Different Standard Speech and Audio Codecs," *IETE Journal of Research*, vol. 58, no. 2, pp. 121–129, 2012.

[8] L. Fernández Gallardo, S. Möller, and J. G. Beerends, "Predicting Automatic Speech Recognition Performance over Communication Channels from Instrumental Speech Quality and Intelligibility Scores," in *submitted to Interspeech 2017*, 2017.

[9] R. P. Lippmann, "Accurate Consonant Perception Without Mid-Frequency Speech Energy," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 66–69, 1996.

[10] J. Rodman, "The Effect of Bandwidth on Speech Intelligibility," 2003, polycom, White Paper.

[11] L. Fernández Gallardo and S. Möller, "Phoneme Intelligibility in Narrowband and in Wideband Channels," in *Annual German Congress on Acoustics (DAGA)*, 2015, pp. 121–124.

[12] M. F. Spiegel, M. J. Altom, M. J. Macchi, and K. L. Wallace, "Comprehensive Assessment of the Telephone Intelligibility of Synthesized and Natural Speech," *Speech Communication*, vol. 9, no. 4, pp. 279–291, 1990.

[13] Y. Teng and R. F. Kubichek, "Speech Intelligibility Evaluation of Low Bit Rate Speech Codecs," in *12th Digital Signal Processing Workshop - 4th Signal Processing Education Workshop*, 2006, pp. 251–256.

[14] E. Jokinen, J. Lecomte, N. Schinkel-Bielefeld, and T. Bäckström, "Intelligibility Evaluation of Speech Coding Standards in Severe Background Noise and Packet Loss Conditions," in *ICASSP*, 2015, pp. 5152–5156.

[15] V. Hazan and A. Simpson, "The Effect of Cue-Enhancement on Consonant Intelligibility in Noise: Speaker and Listener Effects," *Language and Speech*, vol. 43, no. 3, pp. 273–284, 2000.

[16] B. T. Meyer, T. Jürgens, T. Wesker, T. Brand, and B. Kollmeier, "Human Phoneme Recognition Depending on Speech-Intrinsic Variability," *The Journal of the Acoustical Society of America*, vol. 128, no. 5, pp. 3126–3141, 2010.

[17] J. G. Beerends, R. A. van Buuren, J. van Vugt, and J. A. Verhave, "Objective Speech Intelligibility Measurement on the Basis of Natural Speech in Combination with Perceptual Modeling," *Journal of the Audio Engineering Society*, vol. 57, no. 5, pp. 299–308, 2009.

[18] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullman, J. Pomy, and M. Keyhl, "Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part II – Perceptual Model," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 385–402, 2013.

[19] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullman, J. Pomy, and M. Keyhl, "Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part I – Temporal Alignment," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.

[20] C. Hoene, J. M. Valin, K. Vos, and J. Skoglund, "Summary of Opus Listening Test Results," Internet Engineering Task Force (IETF), Tech. Rep., 2011.

[21] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ, USA: Prentice Hall, 1993, ch. 2. The Speech Signal: Production, Perception, and Acoustic-Phonetic Characterization, pp. 11–37.

[22] A. Jongman, R. Wayland, and S. Wong, "Acoustic Characteristics of English Fricatives," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1252–1263, 2000.