Phoneme Intelligibility in Narrowband and in Wideband Channels

Laura Fernández Gallardo, Sebastian Möller
Quality and Usability Lab, Telekom Innovation Labs, TU Berlin, Deutschland
Email: (laura.fernandez-gallardo|sebastian.moeller)@telekom.de

Abstract

With the advent of wideband technologies (50–7,000 Hz), higher transmitted signal quality can be achieved in contrast to traditional narrowband communications (300–3,400 Hz). It is commonly acknowledged that the low frequencies incorporated contribute to increased naturalness, presence, and comfort, whereas the high frequency extension facilitates fricative differentiation. However, no formal intelligibility test to demonstrate this last fact is known to the authors. The present investigation addresses the effects of bandwidth limitation on phoneme intelligibility. Listeners were asked to discriminate among differently bandwidth-filtered and coded-decoded logatome segments, selecting the logatome heard from a list of given options. The results show a statistically significantly better differentiation between /s/ and /f/ in wideband compared to narrowband, not manifest for other tested phonemes.

Introduction

The human capability to recognize speech can be evaluated by conducting intelligibility tests, which measure the identifiability of words or monosyllables previously altered depending on the test objectives. While many intelligibility tests have examined the intelligibility of synthetic speech [1, 2], other tests use natural speech and have focused on the effects of background noise [3, 4, 5, 6]. However, the influence of transmission channels of different bandwidths on phoneme intelligibility has been little studied.

An increasing number of services nowadays is able to deliver narrowband (NB, 300–3,400 Hz) and wideband (WB, 50–7,000 Hz) speech. The added frequency components in the extended bandwidth account for increased naturalness, speech quality, and more accurate speaker recognition and automatic speech recognition. Better word intelligibility can also be attributed to the bandwidth extension [7]. However, to the best of our knowledge, no formal intelligibility test has shown this advantage of WB over NB. This paper presents an intelligibility test from logatomes, where the contribution of different consonant and vowel sounds to intelligibility is analysed for both speech bandwidths.

Previous work

Most intelligibility tests employ word or sentence material, as they represent realistic conditions with which humans are confronted. In these tests, however, the effects of context and predictability of sentences and of word probability in a language must be carefully considered [8]. The Diagnostic Rhyme Test and the Modified Rhyme Test, typically used up to the 1980s, study only the confusability of initial and final consonants and have been criticised for not being sensitive and for overestimating the intelligibility. They use closed sets, in which the listeners are asked to select one out of two or out of six alternatives, and their results may be biased towards more frequent words in a language. Differently, other tests employ non-sense combinations of vowels and consonants [9, 4, 6]. The CLuster-IDentification test (test) has been proposed to overcome the problem of rhyme tests [10]. The intelligibility is evaluated in this case by employing sequences of consonants and vowels in an open-set test. The audio stimuli (monosyllabic words with or without meaning) are generated from combinatory matrices which consider the phonotactic relations and constraints in a language—different consonants and vowels can be combined in the same stimulus. A good overview and discussion of appropriate intelligibility tests in the speech synthesis community is given in [11].

Background noise is considered one of the main factors affecting speech intelligibility of logatomes (nonsense syllables in the form Vowel-Consonant-Vowel (VCV) or CVC) and of CV syllables [3, 4, 5, 6]. In quiet conditions, the study in [5] (in English language) showed that the most confused consonants were the fricatives /ð/-/θ/, /ð/-/v/, /ʒ/-/ʃ/, /θ/-/f/ and, to a lesser extent, the stop sounds /p/-/b/.

Some intelligibility tests have been conducted examining the effects of signal bandwidth to contribute to the research on human perception and hearing. It was found in [9] that the acoustic cues in the high-frequency regions (above 4 kHz) were redundant with those in the mid-frequency regions (0.8–4 kHz), employing CVC logatomes. The differences in intelligibility between NB and broadband speech were analysed in [8] from sentence stimuli, where it was reported that the NB speech required a higher sound level to meet the same speech reception threshold as in broadband. The recent analysis in [12] assessed the importance of various frequency regions for intelligibility from sentences and from phonetically balanced words. The authors found that the frequencies around 1,370 Hz and 2,500 Hz contributed to speech intelligibility more than other examined bands (up to 9,500 Hz).

The intelligibility of natural and of synthesized speech in telephony has been studied in [1] from VCV and CV logatomes employing an open-response test. For natural speech only, the results indicated an intelligibility decrease of around 5% when comparing a headphone condition (clean channel) and a handset condition (telephone channel). The same conditions were also considered in [2], employing meaningful segments (surnames and addresses) uttered by only one speaker. Only marginal effects of bandwidth reduction on initial and final consonant intelligibility were reported.

Audio Material and Transmission Channels

Logatomes in the form VCV were employed as stimuli in order to study the effects of varying the vowel or the consonant sound only. Although it would have been desirable to conduct auditory tests with meaningful well-known words, it was not possible to find or to create a dataset of monosyllabic words differing in only one phoneme and not including different consonants or different vowels in one sample.

The Oldenburg Logatome Corpus (OLLO) [6] contained logatomes suitable for this study. The consonants of the logatomes were selected considering previous phoneme confusions from monosyllabic and bisyllabic rhyme tests in English and in German. The OLLO data, originally intended for speech intelligibility studies under the effects of masking noise [6], was recorded in sound-insulated audiometry rooms (reverberation time $\approx 0.25s$) with a studio-quality condenser microphone. Because the speech data in this dataset is clean, unprocessed, and with sample frequency of 16 kHz, it was possible to transmit the signals through NB and WB communication channels. The different types of distortions applied to the signals could then be controlled in this manner.

The consonants selected for this study were:

- The fricatives /f/, /s/, /v/, and /ʃ/
- The nasals /m/ and /n/
- The stop sounds /b/ and /p/

These eight consonants were embedded in the syllables /afɑ/, /asɑ/, /avɑ/, /aʃɑ/, /amɑ/, /anɑ/, /abɑ/, and /apɑ/ (in German: affa, assa, awwa, ascha, amma, anna, abba, and appa), and the stop consonant /p/ was also embedded in the four syllables /ɛpə/, /ɪpi/, /ɔpo/, and /ʊpu/ (in German: eppe, ippi, oppo, and uppu). These twelve logatomes were selected from the OLLO set of ten German speakers (five males and five females) of standard High German dialect.

The transmission channels studied were a NB channel with the codec AMR-NB operating at 4.75 kbit/s and a WB channel with the codec G.722 at 64 kbit/s. The process for speech distortion was applied as follows. First, the speech was level-equalised 26 dB below the overload of the digital system (-26 dBov), a characteristic level of telephone channels, using the voltmeter algorithm of the International Telecommunication Union (ITU-T) Recommendation P.56. For the NB channel, the signal was downsampled to 8 kHz via an anti-aliasing low-pass FIR filter and then band-passed according to the ITU-T Recommendation G.712 standard implementation. This filter has a flat band-pass response over 300–3,400 Hz approximately. Then, the AMR-NB codec was simulated by employing the tools provided by the European Telecommunications Standards Institute (ETSI), specified in *ETSI EN 301 704*. Finally, the coded-decoded speech was again level-equalised to -26 dBov.

For the WB channel, the original signals, sampled at 16 kHz, were level-equalised to -26 dBov and then band-pass-filtered complying with ITU-T Recommendation P.341. The response frequency of this filter is flat in the range 50–7,000 Hz, approximately. Next, the G.722 encoding and decoding processes were simulated by using standard ITU tools. After applying the codec, the signals were again level-equalised to -26 dBov.

Intelligibility Test

The main objectives of the intelligibility test were to find differences between NB and WB in the detection of particular phonemes, and to quantify the possible improvement of intelligibility with WB communications over NB. A closed-set response format was adopted. The total of stimuli heard by each listener was 192, resulting from twelve logatomes, two channel conditions, and eight repetitions from different speakers in question speaking style. The speakers were randomly selected from the OLLO set of ten speakers with gender balance.

30 listeners (16 males and 14 females) with German as their mother tongue took part in the listening test. Their age ranged from 20 to 35 years, with a mean of 27.0 years. A simple program for presenting the list of the twelve possible logatomes and for logging the listeners' answers was written in Java. The listeners were instructed to click on the label corresponding to the logatome heard. No time constraints were imposed, yet they could listen to each stimulus only once. The sessions were held in a quiet office room and the test administered with a computer with a high-quality soundcard and Shure SRH240 headphones (frequency range 20–20,000 Hz) with diotic listening. Each of the individual listening test sessions took about 12 min to complete, including one break.

The average accuracy reached by the group of listeners was 92.80% with a standard deviation of 25.85%. Figure 1 presents the listeners' accuracies recognizing each logatome in NB and in WB. Despite the high and almost saturated accuracy, caused by the relatively non-severe distortions of the stimuli, significant differences between the two bandwidths could be obtained. The McNemar's test indicated that the difference between the NB and the WB accuracies is statistically significant ($p < 0.001$) for /afɑ/, for /asɑ/, and considering all logatomes pooled. The intelligibility accuracy was improved from NB to WB in every case, except for the logatome /ʊpu/. 22 out of 30 listeners reported they had had difficulties distinguish-

Table 1: Confusion matrix among logatomes in narrowband. The shaded cells correspond to the matrix diagonal.

	/afɑ/	/asɑ/	/abɑ/	/apɑ/	/avɑ/	/amɑ/	/anɑ/	/aʃɑ/	/ɛpə/	/ɪpi/	/ɔpo/	/ʊpu/
/afɑ/	.76	.21		.02	.01							
/asɑ/	.13	.83			.01			.02				
/abɑ/			.84	.10	.05							
/apɑ/			.06	.93								
/avɑ/		.02			.97							
/amɑ/						.95	.05					
/anɑ/							1.00					
/aʃɑ/	.01	.10						.89				
/ɛpə/									1.00			
/ɪpi/									.05	.95		
/ɔpo/											.92	.08
/ʊpu/											.16	.84

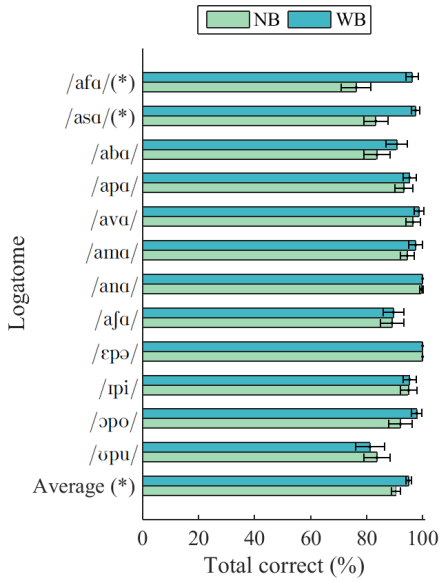

Figure 1: Accuracies detecting logatomes in NB and in WB. Significant statistical differences between NB and WB stimuli with $p < 0.001$ are indicated with (*) for the corresponding logatomes.

ing between /ɔpo/ and /ʊpu/ because of the ambiguous realisations of some speakers, which were perceived as /ɔpu/.

The work in [9], employing CVC logatomes, showed the recognition accuracy of different consonants employing high-pass filters with increasing cut-off frequencies. The accuracy detecting /s/ and /f/ decreased only with a cut-off frequency above 8 kHz and 10 kHz, respectively. Other fricatives and affricates such as /θ,ʃ,tʃ,dʒ/, with frication energy concentrated in lower frequencies, offered an earlier drop in detection performance, when they were high-pass filtered at 4 kHz. Stop sounds such as /p/ and /b/ were less affected by the high-pass filtering. In the experiments of this section, the relevance of the higher frequencies for the recognition of /f/ and /s/, not manifested for other phonemes tested, is confirmed. The inclusion of the high frequencies 3.4–7 kHz in the speech bandwidth enables a significantly better discrimination between /f/ and /s/. The range 50–300 Hz is also included in WB with respect to NB, although this low frequency range presumably offers little benefit in comparison to the high frequencies.

The confusion matrices among logatomes are shown in Tables 1 and 2 for NB and for WB, respectively. Rows denote presented logatomes and the numbers are normalised to the interval 0–1. The confusions with a normalized value lower than 0.01 were omitted from the tables. The greatest confusion can be observed in NB between the logatomes /afɑ/ and /asɑ/ reciprocally, /s/ being better detected than /f/. When switching to the enhanced bandwidth, the total number of errors with each of these logatomes was reduced from 82 in NB to 13 in WB, out of the 240 logatome presentations in each bandwidth. This error reduction is of approximately factor 6. The decrease of confusions was hypothesised, since /s/ and /f/ have similar spectral characteristics in NB but different in WB. Most of the spectral energy of /s/ is concentrated in the higher frequency range incorporated by WB whereas the energy in the /f/ spectrum is more uniformly distributed [13, 7]. Some confusion was also produced between the logatomes /ɔpo/ and /ʊpu/, due to the doubtful /ɔpu/ utterances. The numbers of errors for these logatomes were not substantially reduced in WB with respect to NB.

A recent study conducted by the creators of the OLLO database examined the human speech intelligibility from logatomes [6]. They were sampled at 16 kHz and stationary noise was introduced, but no channel transmissions or bandwidth filters were involved. Considering their results for all speaking styles (normal, question, slow, fast, loud, and soft) and only the consonants also studied in our work, high reciprocal confusions were found between the phonemes: /p/-/b/, /b/-/v/, /f/-/v/, and /n/-/m/. The confusion between /f/ and /s/ was only higher when the presented stimulus was /f/, yet not predominant over the rest of confusions. According to the analyses in [3] and in [5], white noise did not cause confusion between /f/ and /s/ as high as for other phonemes either, for speech band-limited to 200–6,500 Hz.

Table 2: Confusion matrix among logatomes in wideband. The shaded cells correspond to the matrix diagonal.

	/afɑ/	/asɑ/	/abɑ/	/apɑ/	/avɑ/	/amɑ/	/anɑ/	/aʃɑ/	/ɛpə/	/ɪpi/	/ɔpo/	/ʊpu/
/afɑ/	.96	.03										
/asɑ/	.02	.98										
/abɑ/			.91	.07	.02							
/apɑ/			.03	.95								
/avɑ/					.99							
/amɑ/						.97	.02					
/anɑ/							1.00					
/aʃɑ/		.10						.90				
/ɛpə/									1.00			
/ɪpi/									.05	.95		
/ɔpo/											.98	.02
/ʊpu/											.19	.81

Conclusions

An auditory test was conducted in this work to test the effects of NB and WB communication channels on the human speech intelligibility. A closed-set response format was employed, where logatome stimuli in NB and in WB were presented to the listeners.

Our results show that the inclusion of a wider range of frequencies in the speech heard benefits the intelligibility of phonemes, particularly of those with energy concentrated on the higher frequencies, such as the fricatives /f/ and /s/. The logatomes /afɑ/ and /asɑ/ are significantly more intelligible in WB than in NB. /afɑ/ is six times more confusable with /asɑ/ in NB compared to WB, due to the similarities of the NB spectra of the fricatives /f/ and /s/.

References

[1] C. Delogu, A. Paoloni, P. Ridolfi, and K. Vagges, "Intelligibility of Speech Produced by Text-to-Speech Systems in Good and Telephonic Conditions," *Acta Acustica united with Acustica*, vol. 3, no. 1, pp. 89–96, 1995.

[2] M. Balestri, E. Foti, L. Nebbia, M. Oreglia, P. L. Salza, and S. Sandri, "Comparison of Natural and Synthetic Speech Intelligibility for a Reverse Telephone Directory Service," in *International Conference on Spoken Language Processing (ICSLP)*, vol. 1, 1992, pp. 559–562.

[3] G. A. Miller and P. E. Nicely, "An Analysis of Perceptual Confusions Among Some English Consonants," *The Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, 1955.

[4] 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.

[5] S. A. Phatak, A. Lovitt, and J. B. Allen, "Consonant Confusions in White Noise," *The Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. 1220–1233, 2008.

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

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

[8] G. S. Stickney and P. F. Assmann, "Acoustic and Linguistic Factors in the Perception of Bandpass-Filtered Speech," *The Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1157–1165, 2001.

[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] U. Jekosch, "The Cluster-Identification Test," in *Multilingual speech input/output assessment, methodology and standardisation*, E. P. . (SAM), Ed. University College London, London. Internal report II.e, Final report, Year three: 1.III.91-28.II.1992, 1992.

[11] R. van Bezooijen and V. van Heuven, "Assessment of Synthesis Systems," in *Handbook of Standards and Resources for Spoke Language Systems*, D. Gibbon, R. Moore, and R. Winski, Eds. New York, NY, USA: Walter de Gruyter, 1997, pp. 481–563.

[12] E. W. Healy, S. E. Yoho, and F. Apoux, "Band Importance for Sentences and Words Reexamined," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 463–473, 2013.

[13] 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.