Duration threshold for identifying speech samples for different phonemes

HENDRIK HUSSTEDT[1,*], SIMONE WOLLERMANN[1,2], DANIEL BANK[1,2], MARIO SCHINNERL[1,2], MARLITT FRENZ[1] AND JÜRGEN TCHORZ[3]

[1] *German Institute of Hearing Aids, Lübeck, Germany*

[2] *Universität Lübeck, Lübeck, Germany*

[3] *Technische Hochschule Lübeck, Lübeck, Germany*

The identification or classification of acoustic objects is important to decide in which way a sound needs to be interpreted and to rate its importance or relevance. In recent studies, it has been shown that the minimal duration of a sound, which is required for a correct identification, could be a useful audiological parameter, e.g. providing information about the hearing ability of a person. In this work, we want to investigate which cues are used by humans to classify a sound correctly as speech. For this purpose, the duration thresholds for the identification of speech samples starting with different phonemes are analyzed for elderly listeners with normal and impaired hearing. To this end, a two-alternative forced choice (2-AFC) method was used, where, as an alternative to speech, a noise signal with a matched frequency spectrum was presented. In contrast to previous studies, there were no frequency cues available and we found no correlation to the pure tone average (PTA) or speech understanding in noise. As one main conclusion, the results suggest that humans primarily exploit the temporal envelope (ENV) rather than the temporal fine structure (TFS) for the identification of short speech samples above hearing threshold and without frequency cues.

INTRODUCTION

In daily life, the identification or classification of sounds enables the construction of an acoustic scenery with certain objects in our brain. It allows us to decide in which way a sound needs to be interpreted, e.g., the sound of a car has other features than a speech signal. Moreover, in certain listening situations, the identification of sounds is required to rate the importance of the corresponding acoustic object, e.g., speech in a classroom, and car sounds in a traffic situation are of high importance. Consequently, the ability to identify sounds, which is sometimes referred to as auditory gnosis, is an important ability of our acoustic perception (Akelaitis, 1944; Hirsh and Watson, 1996). An impairment of this ability can have serious consequences. If cognitive disabilities are the reason, we speak about auditory agnosia (Pietro *et al.*, 2016). Besides, from an audiological point of view, it is also interesting to investigate how a limited hearing ability can impair the identification of sounds. As depicted by

*Corresponding author: h.husstedt@dhi-online.de

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

Ballas (1993), Gygi *et al.* (2004), and McDermott and Simoncelli (2011), frequency and temporal information are used for the identification, which are both impaired for people with hearing loss (Moore, 1984). For an audiological evaluation, a test paradigm for the identification task has to be defined. To this end, the minimal duration of a sound required for a correct identification can be measured, which was applied in recent studies (Gray, 1942; Bank *et al.*, 2019; Budathoki *et al.*, 2019). These studies report duration thresholds for speech in the range of 20-40 ms, which is remarkably short compared to technical algorithms, e.g., used for the automatic selection of hearing aid programs (Husstedt *et al.*, 2018). Both studies used an alternative forced choice (AFC) procedure with multiple sound samples of different classes, e.g., speech, music, or animal sounds. One drawback of this approach is that the results strongly depend on the sound samples and sound classes chosen (Husstedt *et al.*, 2019). Therefore, another study used the method of adjustment where single sound files could be analyzed independently. However, this method has a high variation due to individual ratings (Husstedt *et al.*, 2019). In this work, we present another measurement procedure to determine the minimum duration of a speech sound required for a correct identification. For this purpose, a 2-AFC task was defined where speech has to be correctly distinguished from a noise signal with matched frequency spectrum. The goal is to investigate which temporal cues are used by humans to classify a sound correctly as speech, and how the perception of those cues is impaired by a hearing loss. To this end, the duration thresholds for the correct identification of speech samples starting with different phonemes are analyzed for a population of 30 people with an age between 60-85 years and a pure tone average (PTA) between 0-80 dB.

MATERIAL AND METHODS

Sound samples

As sound samples, ten German words spoken by a male speaker have been recorded for the test. Different initial phonemes have been selected to investigate their influence on the duration threshold (see Tab. 1).

Abteilung	**Ak**tivitäten	**Al**ternativen	**An**gebote	**Ar**beitsplätze
Bakterien	**Ka**ndidaten	**La**ndwirtschaft	**Na**chmittage	**Ra**hmenbedingungen

Table 1: Ten German words used for the tests spoken by one male speaker.

Five words start with the consonants “B”, “K”, “L”, “N”, and “R” followed by the same vowel “A”. Thus, if there is any effect of the second phoneme, this is almost constant for all of these five words. In contrast, the other five words start with the vowel ”A” followed by the five consonants mentioned so that the effect of the second phoneme can be investigated. For the cutting of the words out of the recorded data, the moving root mean square (RMS) with a window size of 10 ms has been computed,

and the point where the sound pressure level first exceeds a level of 45 dB was set as the starting point. For all words, the starting point is kept constant and the end point is varied to provide sound samples with different durations. Moreover, for each duration of a speech sample, an individual noise signal was generated by randomizing the phase information. Thus, the absolute value of the frequency spectrum is equal for both the speech and the corresponding noise sample. Before the presentation a fade in and fade out of 0.5 ms was applied to all signals. Furthermore, since previous studies indicate that a logarithmic representation of the duration is preferable, the duration of the samples is varied in steps of 1 dB and the absolute duration is provided in Decibel relative to 1 ms (e.g., as illustrated in Fig. 1, 36 dB rel 1 ms corresponds to approx. 63 ms).

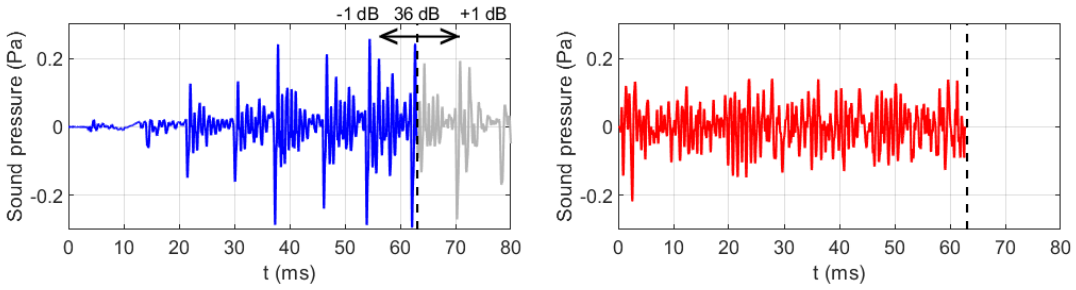

Fig. 1: Visualization of a speech sample (left) and the corresponding noise signal (right). For each duration of the speech sample, the noise signal was generated by randomizing the phase information so that the absolute value of the frequency spectrum is preserved.

Measurement procedure

A touch screen was used for the interaction with the test persons, and the sounds were presented with Sennheiser HDA 280 headphones. First, a dialog was shown where the test persons could listen to different sound samples of different duration, and could adjust the sound pressure level for comfortable loudness. For the test, this selected sound pressure level was applied to each speech and noise sample.

After adjusting the presentation level, the 2-AFC-procedure started where the speech and noise signal were presented in randomized order, and the test person needed to answer which signal was speech (see Fig. 2 a, b). The duration of the samples was adaptively changed with the weighted up-down method according to Kaernbach (1991). The starting point was always 46 dB rel 1 ms (approx. 200 ms), and the end of the test was reached after 12 reversal points (see Fig. 2 c). Until the fourth reversal point, step sizes of 2 dB down and 6 dB up, and afterwards 1 dB down and 3 dB up were used. As result, the average of the last four reversal points was computed and saved as duration threshold.

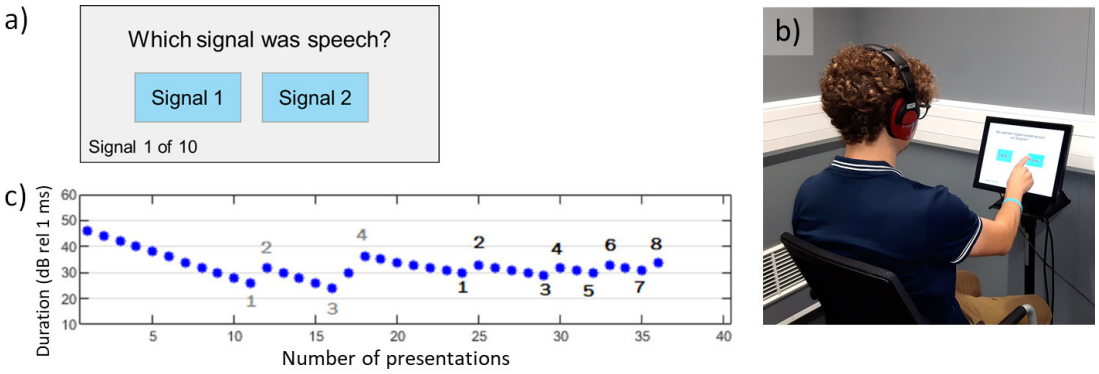

Fig. 2: Visualization of the 2-AFC procedure applied: a) Dialog presented on the touch screen; b) Photograph of a test person typing in an answer; c) Example result for the adaptively changed duration of the samples with the weighted up-down method according to Kaernbach (1991).

Study design

Overall, 30 elderly people with an age between 60-85 years and a pure tone average (PTA) between 0-80 dB were tested. We have defined the PTA as the average of the hearing thresholds for the frequencies 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz. To distinguish effects of hearing loss from age related effects, a population was selected, which shows no significant correlation ($p = 0.26$) between age and hearing loss (see Fig. 3).

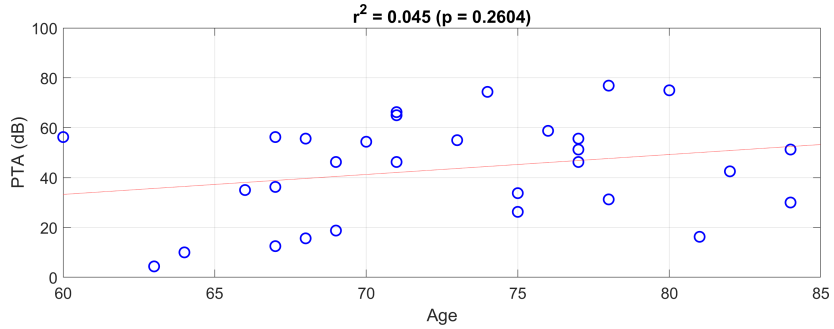

Fig. 3: Scatter plot of the pure tone average (PTA) against age of the 30 test persons.

During the appointment various audiological assessments were performed in the following order: otoscopy, pure tone audiometry, Montreal cognitive assessment (MOCA, Nasreddine *et al.*, 2005), speech in noise with the German sentence test GÖSA (Kollmeier and Wesselkamp, 1997), and finally the measurement of the duration thresholds as depicted in Fig. 2 for the ten German words listed in Tab. 1. The order of the words was chosen according to a Latin square so that after ten test

persons every word has been presented at each position. Furthermore, the order of the columns and rows of the Latin square was randomized for every ten test persons so that the order of the words and their position relative to each other was also randomized.

RESULTS

Initial phonemes

In Fig. 4, the duration thresholds for all words are depicted in a boxplot, and significant differences are indicated with stars. Since the Lilliefors test suggests a normal distribution only for some of the words, the Friedman test including the Tukey-Kramer method was used for a comparison. There are no significant differences between all of the words starting with the same phoneme "A". On the contrary, various significant differences can be observed for the words starting with different phonemes.

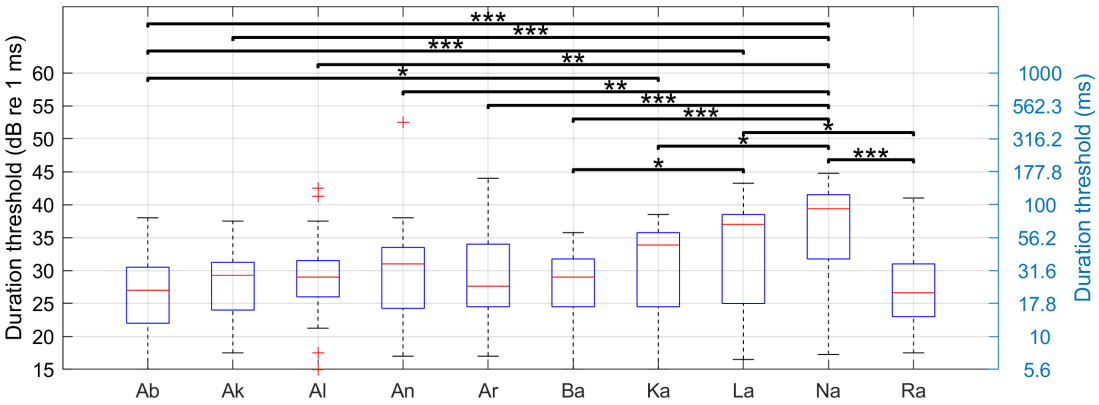

Fig. 4: Boxplot of the duration threshold for all words. For the comparison, the Friedman test including the Tukey-Kramer method was performed. The stars indicate significant differences where "*" corresponds to $p < 0.05$, "**" to $p < 0.01$, and "***" to $p < 0.001$.

Duration thresholds against PTA

In Fig. 5, the duration thresholds of all words are depicted in scatter plots against PTA. For each plot, the Spearman correlation coefficient was computed and it was tested for significance. Since ten hypotheses are tested, the level of significance shall be corrected according to the Bonferroni method ($\tilde{\alpha} = \alpha/10 = 0.05/10$). With this correction, there are no significant correlations between the duration threshold of a word and the PTA. The same holds, if we average the duration threshold over all words.

Speech in noise and cognition

As expected, the results of the speech in noise test (GÖSA) show a high correlation to the PTA whereas there is no correlation to the duration thresholds. Furthermore,

we found a significant correlation between the results of the MOCA and the duration thresholds.

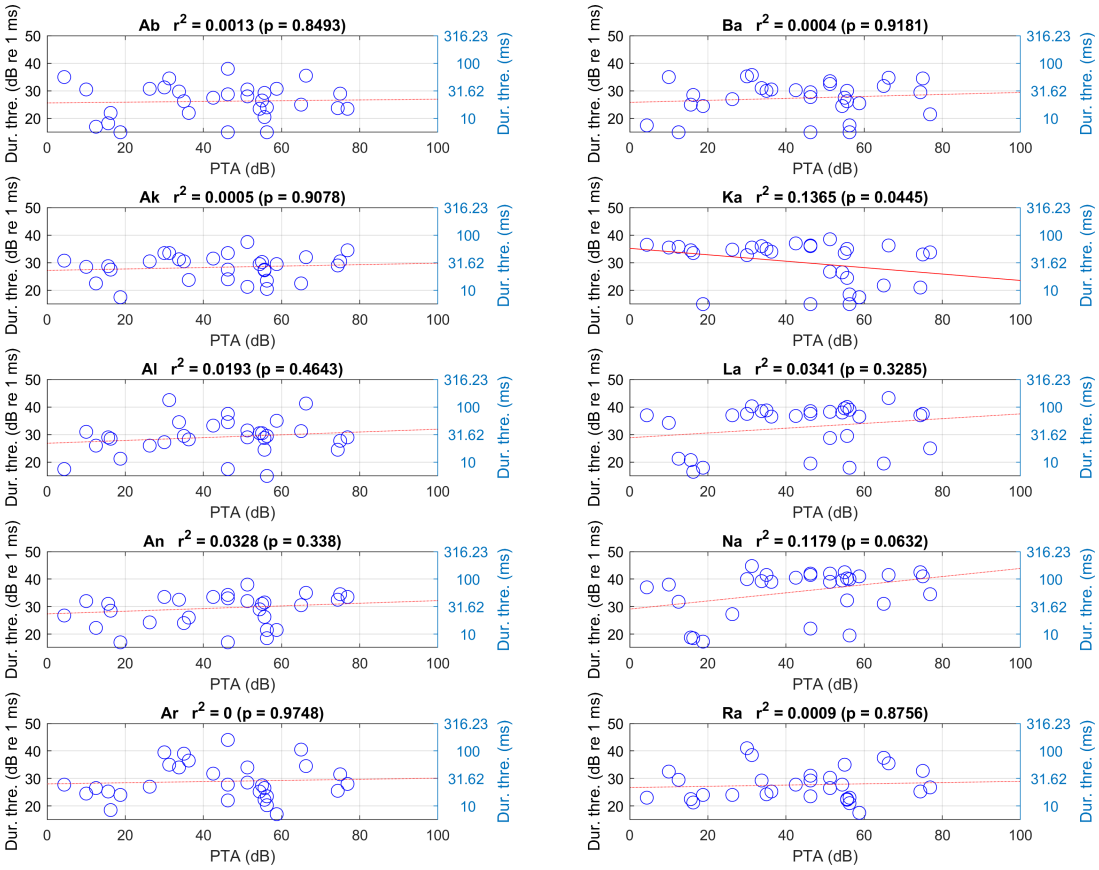

Fig. 5: Scatter plot of the duration threshold against pure tone average (PTA) for all words. Since the α-value is adapted, the p-values indicated are without Bonferroni correction.

DISCUSSION

Initial phonemes

The results clearly demonstrate that different initial phonemes have an influence on the duration threshold. The median is lowest for “Ra” with 21.4 ms and highest for “Na” with 93.1 ms. There are no significant differences between those words starting with the same phoneme “A” followed by different phonemes. This indicates that the second phoneme has a minor or negligible influence on the duration thresholds. A subjective evaluation of the duration of the initial phonemes results in values of approx. 77-120 ms for all words starting with “A”. Since this is longer than the duration thresholds measured (see Fig. 4), this a reasonable explanation for the minor impact of the second phoneme. Except for the word starting with “Ba”, the same holds for all other words

considered. Consequently, in most cases, the initial phoneme is sufficient to correctly detect the signals as speech.

Hearing ability

One main research question of this study is how the hearing ability affects the identification of short speech samples. In contrast to previous studies (Bank *et al.*, 2019; Budathoki *et al.*, 2019), we found no correlation between PTA and duration threshold. As one main difference, in our study there were no frequency cues available due to the matched frequency characteristic of the speech and noise samples. Consequently, one could conclude that a limited frequency resolution of hearing impaired listeners has led to higher duration thresholds in Bank *et al.* (2019) and Budathoki *et al.* (2019).

Without frequency cues, people have to only rely on temporal cues such as the temporal envelope (ENV) and the temporal fine structure (TFS). The results of Lorenzi *et al.* (2006) indicate that hearing impaired people can exploit the information provided by the ENV in the same way as people without hearing loss whereas hearing impaired people show a greatly reduced ability to use information provided by the TFS. Since we see no effect of the hearing loss, the results suggest that the ENV rather than the TFS is used for the identification of short speech samples, if frequency cues are not available and the samples are presented above hearing threshold.

CONCLUSION

In this work, the duration thresholds for the identification of short speech samples starting with different phonemes are analyzed for elderly listeners with normal and impaired hearing. To this end, a 2-AFC method was used where as alternative to speech a noise signal with matched frequency spectrum was presented. The results show that mostly within the first phoneme, people can correctly identify a signal as speech. Moreover, the initial phoneme has a significant influence on the time required for a correct identification. In contrast to previous studies, there were no frequency cues available and we found no correlation to the pure tone average (PTA) or speech understanding in noise. As one main conclusion, the results suggest that humans primarily exploit the temporal envelope (ENV) rather than the temporal fine structure (TFS) for the identification of short speech samples, which include no frequency cues and which are presented above hearing threshold.

REFERENCES

Akelaitis, A. J. (**1944**). “A Study of Gnosis, Praxis and Language Following Section of the Corpus Callosum and Anterior Commissure,” J. Neurosurg., **1(2)**, 94-102. doi: 10.3171/jns.1944.1.2.0094

Ballas, J. A. (**1993**). “Common factors in the identification of an assortment of brief everyday sounds,” J. Exp. Psychol. Hum. Percept. Perform., **19(2)**, 250-267. doi: 10.1037/0096-1523.19.2.250

Bank, D., Schinnerl, M., Frenz, M., Gassenmeyer, F., and Husstedt, H. (**2019**). "Duration Threshold for Identifying Sound Samples of Elderly Hearing Impaired," The Student Conference of the BioMedTec Science Campus, Lübeck, Mar., 2019

Budathoki , D., Tchorz, J., and O'Beirne , G. (**2019**). "Duration Thresholds for Identifying Different Sound Types," 22. DGA Jahrestagung, Heidelberg, Germany, 2019.

Gray, G. W. (**1942**). "Phonemic Microtomy: The Minimal Duration of Perceptible Speech Sounds," Speech Monogr., **9(1)**, 75-90. doi: 10.1080/03637754209390064

Gygi, B., Kidd, G. R., and Watson, C. S. (**2004**). "Spectral-temporal factors in the identification of environmental sounds," J. Acoust. Soc. Am., **115(3)**, 1252-1265. doi: 10.1121/1.1635840

Hirsh, I. J., and Watson, C. S. (**1996**). "Auditory psychophysics and perception," Ann. Rev. Psychol., **47(1)**, 461-484. DOI: 10.1146/annurev.psych.47.1.461

Husstedt, H., Bank, D., and Schinnerl, M. (**2019**). "Comparison of Two Procedures to Measure the Duration Threshold for Identifying Sound Samples," 22. DGA Jahrestagung, Heidelberg, Germany, 2019.

Husstedt, H., Wollermann, S., and Tchorz, J. (**2018**). "Analysis of the Transition of the Automatic Selection of Hearing Aid Programs," 45th Erlanger Kolloquium, Erlangen, Germany, 2018.

Kaernbach , C. (**1991**). "Simple adaptive testing with the weighted up-down method," Percept. Psychophys., **49(3)**, 227-229. doi: 10.3758/BF03214307

Kollmeier, B., and Wesselkamp, M. (**1997**). "Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment," J. Acoust. Soc. Am., **104(2)**, 2412-2421. doi: 10.1121/1.419624

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. (**2006**). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," Proc. Natl. Acad. Sci. U. S. A., **103(49)**, 18866-18869. doi: 10.1073/pnas.0607364103

McDermott, J. H., and Simoncelli, E. P. (**2011**). "Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis," Neuron, **71(5)**, 926-940. doi: 10.1016/j.neuron.2011.06.032

Moore, C. J. M. (**1984**). "Frequency selectivity and temporal resolution in normal and hearing-impaired listeners," Br. J. Audiol., **19(3)**, 189-201. doi: 10.3109/03005368509078973

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., and Chertkow, H. (**2005**). "The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment," J. Am. Geriatr. Soc., **53**, 695-699. doi: 10.1111/j.1532-5415.2005.53221.x

Pietro, M., Laganaro, and M., Schnider, A. (**2016**). "Auditory agnosia," in *Neuropsychological Research: A Review*, Edited by P. Marien and J. Abutalebi (Psychology Press), chap. 15.