PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Word error and confusion patterns in an audiovisual German matrix sentence test (OLSA)

Gerard LLORACH; Volker HOHMANN

Hörzentrum Oldenburg GmbH, Germany

Medizinische Physik and Cluster of Excellence Hearing4all, Universität Oldenburg, Germany

ABSTRACT

One of the established tests for speech intelligibility in quiet and in noise is the Oldenburg sentence test OLSA. This test evaluates the speech reception threshold (SRT), i.e., the speech level or signal-to-noise ratio that leads to a specified word intelligibility, using phonetically balanced 5-word sentences, where each word is equally intelligible. The standard matrix sentence test is audio-only, although it is known that adding visual information to a speech intelligibility task can change the outcome, e.g., increasing the speech intelligibility, or inducing the McGurk effect. Adding visual cues, i.e., being able to see the speaker, could alter the homogeneity of the words with respect to their intelligibility and the expected outcome of the matrix sentence test. In this work we compare the word-error and confusion patterns of the audio-only trials versus the audio-visual ones. Possible reasons of these effects are discussed.

Keywords: audiovisual, matrix sentence test, OLSA

1. INTRODUCTION

In nowadays hearing clinics, there are several methods for testing the hearing ability of an individual. General features about hearing, such as hearing thresholds and loudness comfort levels, are commonly checked. The capabilities of understanding and hearing speech are crucial for human communication. Thus, speech intelligibility tests are also included.

The matrix-sentence-test [1] uses short sentences to test speech intelligibility. The structure of the sentences is always the same, but the content cannot be predicted. Each sentence contains five words: a person name, a verb, a number, an adjective and an object. The words used in these sentences are chosen to be known and common in the spoken language and to represent the phoneme distribution of the language. There are ten different words for each position in the sentence (ten names, ten verbs...) and they are combined to create sentences. A test commonly contains fifteen to twenty sentences. The advantage of this test is that it uses grammatically correct sentences that cannot be semantically predicted. Additionally, the sentences can be reused in different conditions because they are not possible to remember, as there are too many possible word combinations.

Most current speech intelligibility tests rely only on audio stimuli. Visual information is an important factor for speech intelligibility, as it increases when seeing the face of the speaker [2]. Individuals can lip-read the mouth movements and integrate it with the acoustic signal. Including visual information in matrix-sentence-tests is crucial to understand other factors that affect speech intelligibility. For example, the audiovisual integration and lip-reading capabilities of an individual could be evaluated using an audiovisual matrix-sentence-test. These factors would give a better image of the problems that the individuals with hearing impairments encounter in real situations, e.g. a good lip-reader might not have problems in face-to-face communication, but when using the telephone.

One of the properties of the matrix-sentence-test is that the intelligibility of each word is similar. This is done to achieve a higher precision when measuring speech reception thresholds (SRTs). When this homogeneity between words is achieved, the speech reception curve becomes steeper, leading to higher precision in the SRTs measurements [3]. Including visual stimuli might affect this homogeneity, as some words might be easier to lip-read than others.

Visual speech can also affect the acoustic perception of speech. The McGurk effect affects speech syllable perception by showing incongruent visual speech [4]. Whether this effect still appears in the matrix-sentence-test is difficult to say. For example, it has been found that the McGurk effect does not have an influence when understanding full sentences [5]. Nevertheless, the OLSA test uses rather short sentences and a small set of words. The visual speech could influence how words are identified with others, in comparison to audio-only experiments.

The goal of this work is to understand how visual information in the German matrix-sentence-test (OLSA) changes the word-error and word confusion patterns. Dubbed video recordings of the original speech material were used. 30 young normal hearing listeners were tested with different sensory modalities (audio-only, video-only, audiovisual). Figure 1 shows that seeing the speaker does not affect the intelligibility between words. Thus, the same property of the audio-only OLSA regarding similar speech intelligibility between words is kept.

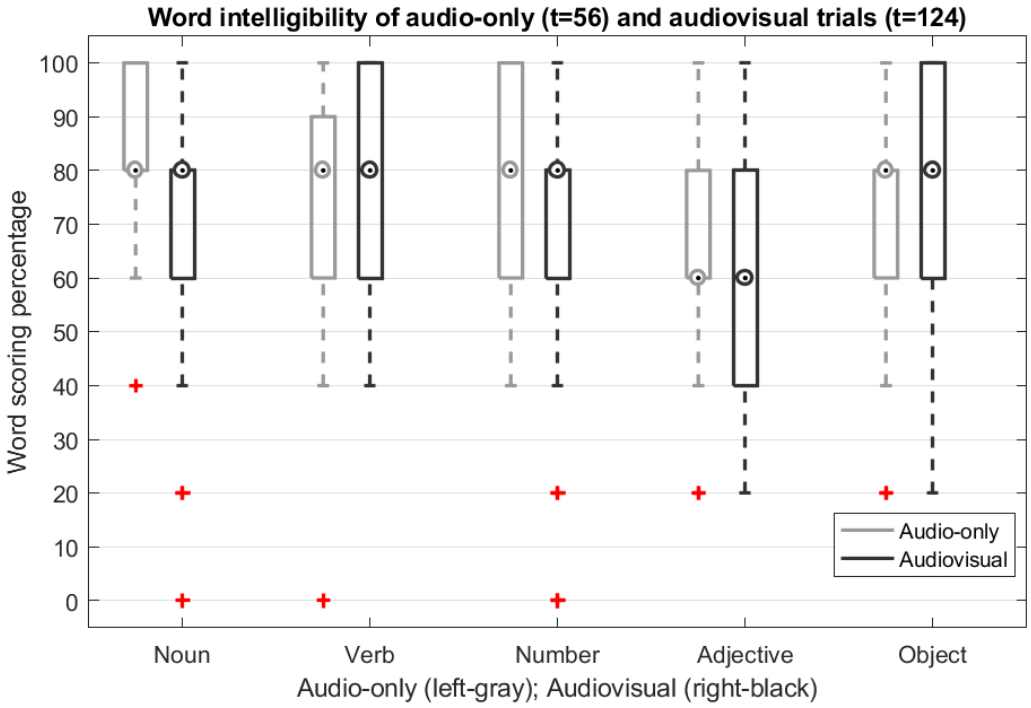

Figure 1. Boxplots of the word intelligibility in audio-only and audiovisual trials. 20 sentences were used per trial. Speech was presented in noise and participants chose the answers from a set of possible words. The speech level was changing to reach an intelligibility of 80%. Only the word intelligibility of the last 5 sentences of each trial was used in this figure.

Table 1 shows some of the effects of adding visual cues on a word level. Some words gained intelligibility when adding visual cues and could be well lip-read when there was no acoustic signal. For example, visual cues increased the speech intelligibility of “Blumen” by 33% and this word could even be lip-read and identified correctly on 75% of the sentences. Nevertheless, some new word-confusions appeared, e.g. “Tassen” was often confused with “Sessel” when visual cues were present (Table 1). These word confusions were most likely due to word similarity or due to containing parts of other word, e.g. “acht” and “acthzehn”. Audiovisual benefit could be attributed to some specific visual speech units (visemes). For example, words starting with “sh”, such as “schwere”, “schöne”, “Stefan”, “schenkt”, “Schuhe” and “Steine”, showed a higher audiovisual benefit and good lip-reading scores. Visual cues helped also to disambiguate between words. For example, “Bilder” and

“Ringe” were often confused in audio-only sentences, but not in audiovisual ones. In summary, the possibility to lip-read the speaker changed the word-error and confusion patterns between words. Some word ambiguities were resolved but some others were introduced. Lip-reading alone could provide speech intelligibility above 65% for certain words, which would invalidate any measurements that target audiovisual speech intelligibility below those levels.

Table 1 – Relevant words that had an increase or decrease in intelligibility when adding visual cues. Best lip-read words and word confusion due to visual cues are also shown.

Audiovisual benefit (> 15%)	**Audiovisual detriment (> 15%)**
bekommt, gibt, verleiht, malt, sieben, fünf, schwere, schöne, Blumen, Bilder, Ringe	Tanja, gewann, drei, achtzehn, kleine, nasse, grüne, teure, Tassen
Best lip-read words (> 65% Speech Intelligibility)	**Word confusions due to visual cues (increase of word confusion by > 15%)**
Ulrich, Wolfgang, Stefan, Thomas, bekommt, kauft, schenkt, verleiht, malt, fünf, Blumen, Schuhe, Steine	Peter-Britta, verleiht-gewann, acht-achtzehn, grüne-rote, Tassen-Sessel

ACKNOWLEDGEMENTS

This work was funded by the European Union’s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 675324 (ENRICH) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Projektnummer 352015383 – SFB 1330 B1.

REFERENCES

1. Wagener K., K ü hnel V. & Kollmeier B. Entwicklung und Evaluation eines Satztests in deutscher Sprache - Teil I: Design des Oldenburger Satztests (in German). (Development and evaluation of a German sentence test - Part I: Design of the Oldenburg sentence test). Z Audiol. 1999; 38, 4 – 15.

2. Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. The journal of the acoustical society of america. 1954; 26(2):212-5.

3. Kollmeier B. Messmethodik, Modellierung, und Verbesserung der Verständlichkeit von Sprache (in German). (Methodology, modeling, and improvement of speech intelligibility measurements). Habilitation thesis. Göttingen: University of Göttingen. 1990.

4. McGurk H, MacDonald J. Hearing lips and seeing voices. Nature. 1976; 264(5588):746.

5. Van Engen KJ, Xie Z, Chandrasekaran B. Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. Attention, Perception, & Psychophysics. 2017; 79(2):396-403.