Speech intelligibility as a function of temporal resolution and the number of stimulation channels for signal processing using the sinusoidal speech model

O. Timms*, S. Allegro[+], V. Kühnel[+], A. von Buol[+], N. Dillier*
***Dept. Otorinolaringology University Hospital Zurich Switzerland; [+]Phonak AG, Stäfa, Switzerland**

Abstract

The number of stimulation channels as well as the temporal and spectral resolution required for 100% speech intelligibility were determined for a sinusoidal speech signal processing system.. German vowel, consonant and sentence tests were processed with an algorithm based on the sinusoidal speech model of Quatieri and McAulay [1]. The algorithm uses a limited number of stimulation channels (1-5) as well as different temporal and spectral resolutions (FFT frame length 128, 256 and 512 points, sampling frequency 22050 Hz). Speech intelligibility tests in quiet were performed with normal hearing native German speaking adults using conventional diagnostic speech perception tests as well as paired comparison consonant tests.

The results show that different numbers of stimulation channels are required for 100% speech recognition in sentences processed with different temporal and spectral resolution. A tendency of increased speech intelligibility was observed by either increasing temporal resolution or increasing the number of stimulation channels. In addition significant learning effects were observed.

Method

A block diagram of the present sinusoidal speech analysis / synthesis system is depicted in figure 1. The most important parts of the original algorithm of Quatieri and McAulay as FFT and peak picking are preserved. However, due to the high complexity of cubic polynomial phase interpolation, frequency tracking and sine wave generation in the original algorithm, an IFFT with a 75% overlap-add is used for synthesis. These simplifications result in a quality reduction of the processed signal. The most remarkable effect is musical or "bubble"-like noise as known from vocoder systems.

Three FFT lengths (128, 256 and 512 points) at a sampling frequency of 22050 Hz corresponding to time resolutions of 5.8, 11.6, and 23.2 ms were tested with 1 to 5 spectral lines.

For peak selection the lowpass filtered smoothed frequency spectrum as proposed by Kates was used [2].

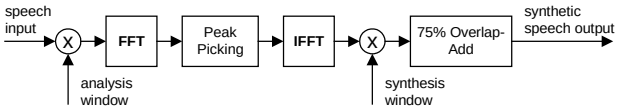

Fig. 1 Sinusoidal analysis/synthesis system block diagram.

Performed Tests

Oldenburg Sentence tests [3] were carried out in quiet with 16 native German speaking adults. The sentences were presented at 65 dB RMS in a sound proof room at a distance of 1.5 m from a Westra Audiometer Box Type LAB - 1001 (playback by a 16 bit PC sound card).

C12 Consonant aCa-Logatome tests were carried out in quiet with 21 native German speaking adults. Logatome syllabics were presented at 65 dB RMS in a sound proof room at a distance of 1.5 m from a Westra Audiometer Box Type LAB - 1001 (playback by a 16 bit PC sound card). The subjects were asked to identify the aCa-syllable from a choice of 12 aCa-syllables on a touch screen (MAC-test [4]).

Vo8 Vowel dV-Logatome tests were carried out with the same test setup as for the consonant tests with 14 native German speaking adults. The subjects were asked to identify the correct dV-syllable from a choice of 8 dV-syllables.

Consonant C12 aCa- Logatome paired comparison tests were carried out in quiet with 24 native German speaking adults. Only those consonant pairs were tested which caused difficulties in the logatome identification tests. Logatome syllables were presented at most comfortable level by Sennheiser headphones Type HD 570 (playback by a 16 bit PC sound card) in a sound proof room. The subjects were asked to judge discrimination certainty of the two presented aCa-logatome syllables.

Results

Figure 2 shows the results of the Oldenburg sentence test. For 100 percent sentence recognition at a temporal resolution of 5.8 ms (128 pt FFT) only one stimulation channel is required. With decreasing temporal resolution the number of spectral channels required for 100 percent recognition increases up to four for a temporal resolution of 23.3 ms (512 pt FFT). The test shows that temporal resolution is critical for speech recognition in quiet. Furthermore, remarkable learning effects were observed during the tests (not depicted).

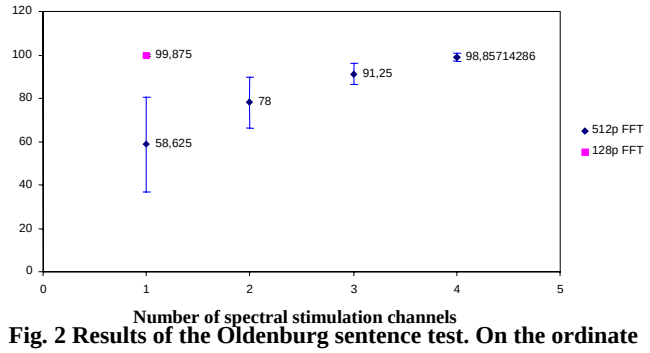

Fig. 2 Results of the Oldenburg sentence test. On the ordinate the score of right answers in percents are given.

In agreement with the results from the sentence tests, the identification scores of both vowels and consonants increase with increasing number of spectral stimulation channels and with increasing temporal resolution.

After being processed, some consonants sound unnatural due to musical noise effects, resulting in low identification scores for the logatome identification tests. In these cases, no dominant logatome confusions are observable in the confusion matrices (figures 3 to 6). For example, for each of the tested temporal resolutions the syllable "aRa" is poorly recognized even with five stimulation channels (see figures 3 and 4). However, in the paired comparison logatome tests the same syllable "aRa" is well separable from the other consonants (figure 7). This might be explained by the unnatural pronunciation of these consonants resulting from the processing artifacts. With training those consonants should become well recognizable.

There are also some consonants and vowels, for example German "S" "A", which are well recognizable independent of the temporal or spectral resolution. For those phonemes the temporal resolution is less significant and the spectral resolution of 174 Hz (128 pt FFT) is sufficient.

The paired comparison tests showed that there are some confusions, for example German "N-M" or "F-P"(figures 7 and 8), where a defined number of stimulation channels is required to reach an acceptable rate of discrimination certainty. This consonants are especially interesting for future investigations.

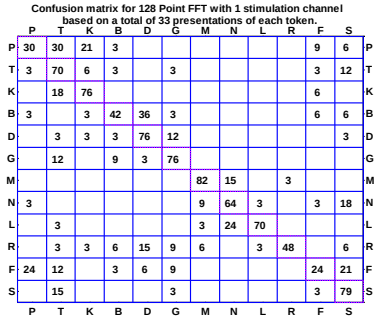

Fig. 3 Confusions matrix for 5.8 ms temporal resolution with 1 stimulation channel (confusions are given in percents).

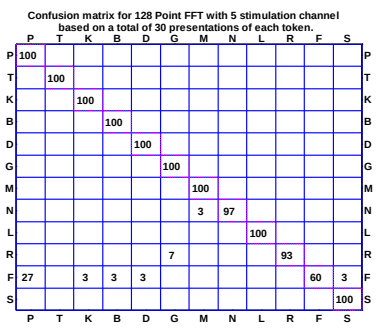

Fig. 4 Confusions matrix for 5.8 ms temporal resolution with 5 stimulation channels (confusions are given in percents).

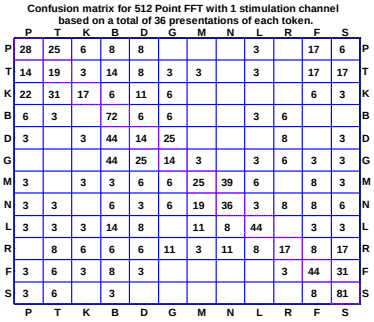

Fig. 5 Confusions matrix for 23.2 ms temporal resolution with 1 stimulation channel (confusions are given in percents).

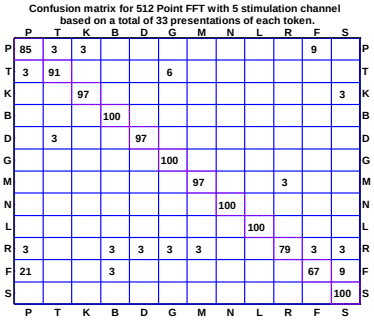

Fig. 6 Confusions matrix for 23.2 ms temporal resolution with 5 stimulation channels (confusions are given in percents).

In the paired comparison tests there are also some consonant confusions which show a decrease of discrimination certainty with growing number of spectral stimulation channels (see e.g. "F-P" in figure 8). These effects can be explained by the uncertainty of the observations.

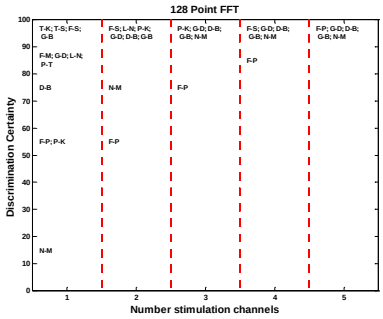

Fig. 7 Results of the Logatome aCa tests for 5.8 ms temporal resolution

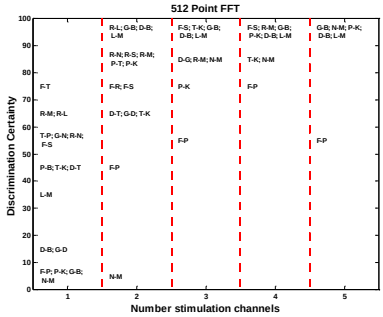

Fig. 8 Results of the Logotome aCa tests for 23.2 ms temporal resolution.

Conclusions

The use of three spectral stimulation channels seems to be sufficient for the recognition and discrimination of German vowels and consonants for each of the three tested temporal and spectral resolutions. With temporal resolutions of 5.8 ms and 11.6 ms, only two spectral stimulation channels are required for good vowel and consonant recognition. With a temporal resolution of 5.8 ms and the use of only one spectral stimulation channel, recognition difficulties occur with the four German consonants "P", "B", "R", and "F" and the three vowels, "I", "Ä", and "Ü".

It was shown that for speech recognition in quiet with a reduced number of spectral stimulation channels, temporal resolution is more important than spectral resolution.

Remarkable learning effects during the sentence- and Logatome tests are can be seen. Paired comparison tests are necessary to get meaningful results.

References

[1] T.F. Quatieri and R.J. McAulay, "Audio signal processing based on sinusoidal analysis / synthesis", chapter 9 of "Applications of Digital signal processing to audio and acoustics", edited by M. Kahrs and K. Brandenburg, Kluwer Academic Publishers, 1998.

[2] J.M. Kates, "Speech Enhancement Based on a Sinusoidal Model," Journal of Speech and Hearing Research, vol. 37 pp. 449-464, 1994.

[3] Wagener, K., Kühnel V. and Kollmeier, B. "Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests", Zeitschrift für Audiologie, 1, 4-15, 1999.

[4] N. Dillier, T. Spillmann, „Deutsche Version der Minimal Auditory capability (MAC) – Test- Batterie" in Kollmeier, B. Moderne Verfahren der Sprachaudiometrie, Median Verlag. Heidelberg, 1992.