1 General Methods

1.1 Stimulus sentences

. . . . . - Sy How sentences were constructed. - Say how cloze probabilites were collected. - Say how 120 and then 360 auditory stimuli are created.

1.2 Speech processing

All 360 sentences used in the experiments in this thesis were first recorded and digitized at 44.1 kHz with 32 bit linear encoding. The quality of auditory stimuli – in chapters 5 through 8 – are manipulated by noise-vocoding (chapter 5 to 8), and speech compression algorithm PSOLA (chapter 7). These speech processing . . . . . .

1.2.1 Noise-vocoding

Noise-vocoding is used to parametrically vary and control the quality of speech signal and a graded manner. It largely removes the spectral details of the speech signal but preserves the temporal and preiodicity cues [1].

Noise-vocoding distorts speech by dividing a speech signal into specific frequency bands corresponding to the number of vocoder channels. The frequency bands are

analogous to the electrodes of cochlear implant [2–4]. The amplitude envelope – fluctuations of amplitude – within each band is extracted and is used to modulate noise of the same bandwidth. It renders vocoded speech harder to understand by replacing the fine structure of the speech signal with noise while preserving the temporal characteristics and periodicity of perceptual cues [1].

If the cut-off frequencies of the bandwidth of the speech signal (i.e., the analysis band) and the bandwidth of the noise do not match then the resulting noise-vocoded speech becomes spectrally shifted [e.g., 5]. The cut-off frequencies of the speech signal and the to-be-modulated noisebands are identical for all the speech stimuli in the current study.

Sentences were noise-vocoded through 1-, 4-, 6- and 8-channels in Experiment 1, 2, and 4 using custom scripts originally written by **Darwin2005**. In Experiment 3, they were vocoded only through 4-channels. The cut-off boundary frequencies were set between 70 Hz and 9000 Hz. Upper and lower bounds for band extraction within each bandwidth of each noise-vocoding condition are shown in Table X.X which follows Greenwood's cochlear frequency position function [6, 7]. Scaling was performed to equate the root-mean-square values of the original undistorted signal and the final noise-vocoded sentences.

Spectrograms of clear speech and noise-vocoded speech (1-, 4-, 6- and 8-channels) for the word 'Aufgabe' are shown in Figure X.X. It shows that with a decrease in the number of noise-vocoding channels, speech signal becomes more and more similar to noise.

1.2.2 Speech compression

Uniform time-compression algorithms like pitch-synchronous overlap-add technique [PSOLA, Charpentier and Stella [8]; Moulines and Charpentier [9]) are used to compress the speech signal and increase the rate of speech. PSOLA analyzes the pitch of an auditory signal, in the time domain of its digital waveform, to set pitch marks and then segments the signal into successive analysis windows centered around those pitch marks. To create synthesized speech, a new set of pitch

marks are calculated and the analysis windows are rearranged. Depending on the time-compression factor, some analysis windows are deleted, and the remaining windows are concatenated by superimposing and averaging the neighboring analysis windows. Hence the resulting speech signal is compressed, i.e., it is perceived to be faster than the original speech [e.g., CITE] The distortion of phonemic properties of speech signals are minimal when accelerating and slowing down within the range of factor 2 or below [9]. In Experiment 3, PSOLA algorithm in Praat software is used to increase the rate of speech by a factor of 0.65 before passing it through 4-channels noise-vocoding.

Spectrograms of clear speech, speeded speech, and noise-vocoded speeded speech for the word 'Aufgabe' are shown in Figure X.X. It shows that .....

1.3 Measurement of language comprehension

References

[1] Stuart Rosen, Andrew Faulkner, and Lucy Wilkinson. "Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants". In: *The Journal of the Acoustical Society of America* 106.6 (Dec. 1999), pp. 3629–3636. URL: http://dx.doi.org/10.1121/1.428215.

[2] R. V. Shannon et al. "Speech Recognition with Primarily Temporal Cues". In: *Science* 270.5234 (Oct. 13, 1995), pp. 303–304. URL: http://dx.doi.org/10.1126/science.270.5234.303.

[3] Philipos C. Loizou, Michael Dorman, and Zhemin Tu. "On the number of channels needed to understand speech". In: *The Journal of the Acoustical Society of America* 106.4 (Oct. 1999), pp. 2097–2103. URL: http://dx.doi.org/10.1121/1.427954.

[4] Robert Shannon, Qian-Jie Fu, and John Galvin Iii. "The number of spectral channels required for speech recognition depends on the difficulty of the listening situation". In: *Acta Oto-Laryngologica* 124.0 (Apr. 1, 2004), pp. 50–54. URL: http://dx.doi.org/10.1080/03655230410017562.

[5] Andrew Faulkner, Stuart Rosen, and Tim Green. "Comparing live to recorded speech in training the perception of spectrally shifted noise-vocoded speech". In: *The Journal of the Acoustical Society of America* 132.4 (2012), EL336–EL342.

[6] Donald D. Greenwood. "A cochlear frequency-position function for several species—29 years later". In: *Journal of the Acoustical Society of America* 87.6 (1990), pp. 2592–2605.

[7] Julia Erb. "The neural dynamics of perceptual adaptation to degraded speech". Doctoral dissertation. Universität Leipzig, 2014, p. 211.

[8] F. J. Charpentier and M. G. Stella. "Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation." In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (1986), pp. 2015–2018.

[9] Eric Moulines and Francis Charpentier. "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones". In: *Speech Communication* 9.1990 (1990), pp. 453–467.