PHONETIC ANALYSIS VS. DIRTY SIGNALS: FIXING THE PARADOX

Tina John[1], Oliver Niebuhr[1], Gerhard Schmidt[2], and Anne Theiß[2]

[1]Institute for Scandinavian studies, Frisian studies and General Linguistics, Kiel University
[2]Digital Signal Processing and System Theory, Kiel University
tjohn@isfas.uni-kiel.de

Abstract: Noisy speech signals and phonetic data analysis: two phrases that rarely occur in the same context. This paper shows that noise cancellation and noise suppression methods can clean noisy speech signals to such an extent that they can be reliably segmented into phones semi-automatically using appropriate tools such as the Munich Automatic Segmentation Tool. This finding brings noisy signals and phonetic data analysis closer together.

1 Introduction

Phonetic and linguistic analyses of speech produced under adverse conditions of physical and cognitive load require recordings in noisy environments, in order to document realistic speech behaviour. An example of such a condition is the communication of drivers and passenger in moving cars, which is characterized by the physical load of driving as well as by the noise of the moving car. Recordings obtained from a microphone in this situation are inevitably impaired by noise. Acoustic measurements, such as the calculation of spectra, cepstra or formants are distorted in these recordings and can thus easily lead to misinterpretations.

To overcome these problems we developed a system that should reproduce pre-recorded acoustic ambiances in a reliable manner (see Jaschke, 2012, pp. 29ff). This is achieved by playing back appropriately equalized noise signals (obtained by so-called "noise-only recordings" in a target environment) via a multitude of loudspeakers in a lab (for details see Sec. 2.1). The equalization is performed such that the (re-) created acoustic conditions in a lab should be as close as possible to the (pre-) recorded real situation. In contrast to real recordings (e.g. in a car) the recordings in the lab can be enhanced significantly by signal processing since reference signals for the noise (i.e. the loudspeaker signals) are available now.

The basic question for such a system is: Is the speech production behaviour the same in the simulated environments as in a real driving situation? Answering this question requires phonetic and linguistic analyses of noisy speech material. According to Jaschke (2012), multi-channel noise cancellation (NC) and noise suppression combined with NC (NC+S) can clean the signal from the noise to a certain extend (SNR improvements of about 30 dB could be achieved). But is this cleaning method good enough to allow reliable analyses of those acoustic parameters that phonetic and linguistic researchers are interested in and are these parameters extractable using already existing algorithms?

The first step in the phonetic analysis of speech data is the segmentation and annotation of the spoken phones, words etc. in the signals. This process is very time consuming, particularly for a larger amount of data. The Munich Automatic Segmentation (MAUS) tool (Kipp et al., 1996) provides a semi-automatic segmentation that might speed up the task. Further, the EMU system (John et al., in press) provides functions for semi-automatic constructions of labelling hierarchies.

This paper will study whether speech signals that were overlaid by noise in an acoustic ambience simulation of a moving car and subsequently underwent NC and NC+S enhancement can be segmented by MAUS with the same reliability as reference signals without any noise. Furthermore, the contributions of the individual processing stages (NC, NC+S) to this reliability will be analysed. The results of the study help to advance signal-

processing methods for the analysis of speech in noisy environments in general and for the improvements of speech enhancement systems such as hands-free or in-car communication systems in particular.

1.1 Ambience simulation and recordings

The acoustic ambience simulation is based on signals of a noisy environment that are first recorded and then played back to listeners via loudspeakers. The signals need to be recorded in real situations. For example, in the case of the present study simulating a moving car required driving a real vehicle at different speeds, on different road surfaces, in different traffic conditions, etc. During these test drives, calibrated in-ear microphones included in a so-called artificial head recorded the environmental noise binaurally and stored the noise as a stereo signal. Figure 1 shows the artificial head that has been used for our recordings. However, before the recorded signals can be used for simulation purposes, they need to be checked for artefacts and prepared for a loop mode. The latter step is to avoid artefacts that can result when switching from the end of the signal to the beginning.

Figure 1 - Test drive in our vehicle (left) with the artificial head (right) including in-ear microphones that record the background noise.

Ambience simulations combine realistic acoustic situations with a controlled, research-friendly environment. In other words, they serve to carry the field into the lab. Such simulations require a soundproof room. The room that has been constructed and furnished with two loudspeakers and two subwoofers around a microphone is displayed in Figure 2.

In the ambience simulation the room acoustics as well as the distance and the angle of the listening person relative to the loudspeakers modifies the signals characteristic at the ear of the person. To ensure less deviation as possible, played back signals are calibrated (cf. Figure 2). That is, the environmental signal is played via loudspeakers to an artificial head or a person wearing in-ear microphones. The microphone signal is compared to the in-ear recordings obtained in the real environment and the loudspeaker signals are modified (iteratively) until both signals are equalized or resemble each other as much as possible.

Prior to recordings, speakers in the soundproof room are familiarized with their task (e.g., reading a story, answering questions, etc.). They are given some time to get used to the lab environment before and after the loudspeaker signals create the ambience simulation. Close-talking microphones record the speakers' utterances and the loudspeakers signals.

1.2 Signal enhancement

The task of the signal enhancement is to clean the signal to such an extent that the background noise inserted by the simulation is significantly reduced but also to keep the speech in a quality good enough for listeners to judge it as good speech signals as well as to keep enough phonetic cues for phonetic analyses.

In our approach two different signal enhancement steps are used. The first step is a noise cancellation method (NC). Here the loudspeakers signal are fed into adaptive filters that try to

estimate the multi-path propagation of the acoustic transmission from the loudspeakers to the microphone that records the speakers. We have used a gradient-based estimation procedure: the so-called normalized least mean square (NLMS) algorithm. After estimating the acoustic channels the loudspeaker signals are convolved with the channel estimation and the resulting signals are subtracted from the microphone signal. Assuming a good match between the true and the estimated transmissions the subtraction can remove the noise without affecting the speech components in the microphone. The second one is the noise suppression method (NS). Here the power spectral density of the remaining noise is estimated in a first step. Afterwards time-variant, frequency selective gain factors are determined according to a so-called Wiener filter. In contrast to the first approach, in this second step (NS) always a compromise between the degree of noise suppression and the introduction of speech distortions has to be found. As Figure 2 shows each signal is noise cancelled and the enhanced signal is further processed by noise suppression. The result then is a noise cancelled and suppressed signal (NC+S).

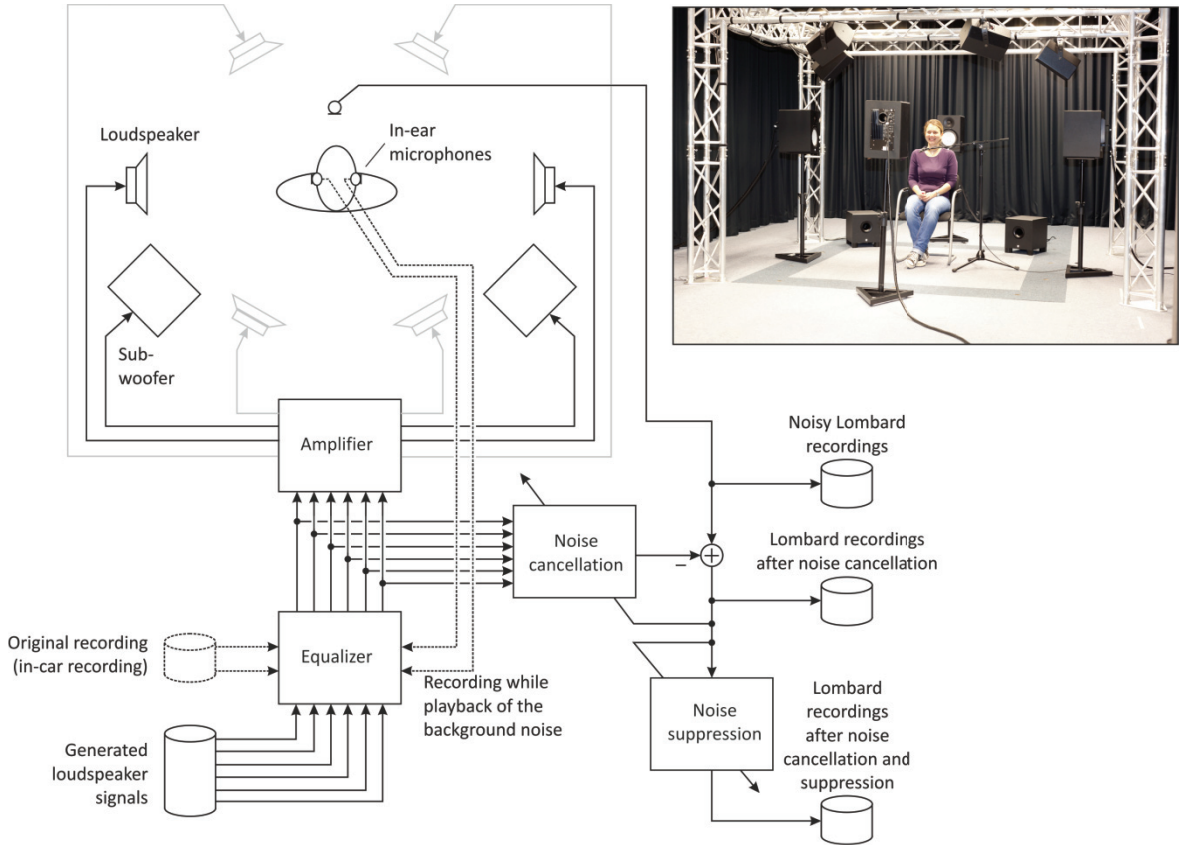

Figure 2 - Ambience simulation setup with equalizing/ calibration of the play back hardware (dotted) and the recording setup with the enhancement methodology

1.3 MAUS Segmentation

The **M**unich **Au**tomatic **S**egmentation (see Kipp et al., 1996; Kipp et al., 1997) is based on a given canonical text, statistical probabilities and a Hidden Markov Model of acoustic correlates of phonetic features. Based on the canonical transcription of the speech signal in SAMPA, pronunciation variants of the words are generated by lexicon lookup and organised in a graph structure. The given graph structure, in turn, constrains an HMM state space that is used in the data driven Viterbi beam search for the time alignment of the segment boundaries. The HMM includes the following acoustic parameters: 12 cepstral coefficients, energy, zero crossing rate and first and second derivative every 10 ms. The segmentation and labelling procedure results in a 10 ms grid, hence imprecise segment boundaries. In the final processing step boundaries are corrected somehow by shifting them to the next zero crossing in the signal. Kipp et al. (1996) has shown that MAUS differs less than 7% from a human

segmentation in terms of average matches between three manual and one automatic segmentation of a large corpus. Hence segmentation is not based on acoustic data only. MAUS formats the outcoming string of phonetic labels and the corresponding segment boundaries as Praat-TextGrids (Boersma, 2002).

2 Methodology

2.1 Material

Since the research question was raised by the work of Jaschke (2012), we used the same speech database as he did. It consists of UN recordings as well as of the audio recordings processed in terms of NC and NC+S. The recordings comprise 10 productions of the story "north wind and the sun" by different speakers. Each speaker read the story in three different noisy in-car environment simulations (IC 1-3: 0, 50, 100 km/h, which corresponds to an increase in noise level), as well as in a reference environment (REF) without any added car noise. The ambience simulation used noise from static traffic without overtaking cars. Speakers (VP1-10) were 5 males and 5 females between the age of 22 and 30.

2.2 Analysis

MAUS was run four times for the speech signals of all four conditions (IC1-3 and REF) with four different input conditions in the form of different accuracies of the canonical text. Input differed as follows: run 1 - canonical transcription of the text; run 2 - individual canonical transcription of the actually spoken words; run 3 - individual transcription extended by assumed major prosodic units based on the orthographic transcription of the narrative; run 4 – further extended by assumed minor prosodic units. The resulting phone segmentations from run 4 were transferred from Praat TextGrids to an EMU database. The annotated database was then semi-automatically extended by word annotations and meta-data (see John, 2012).

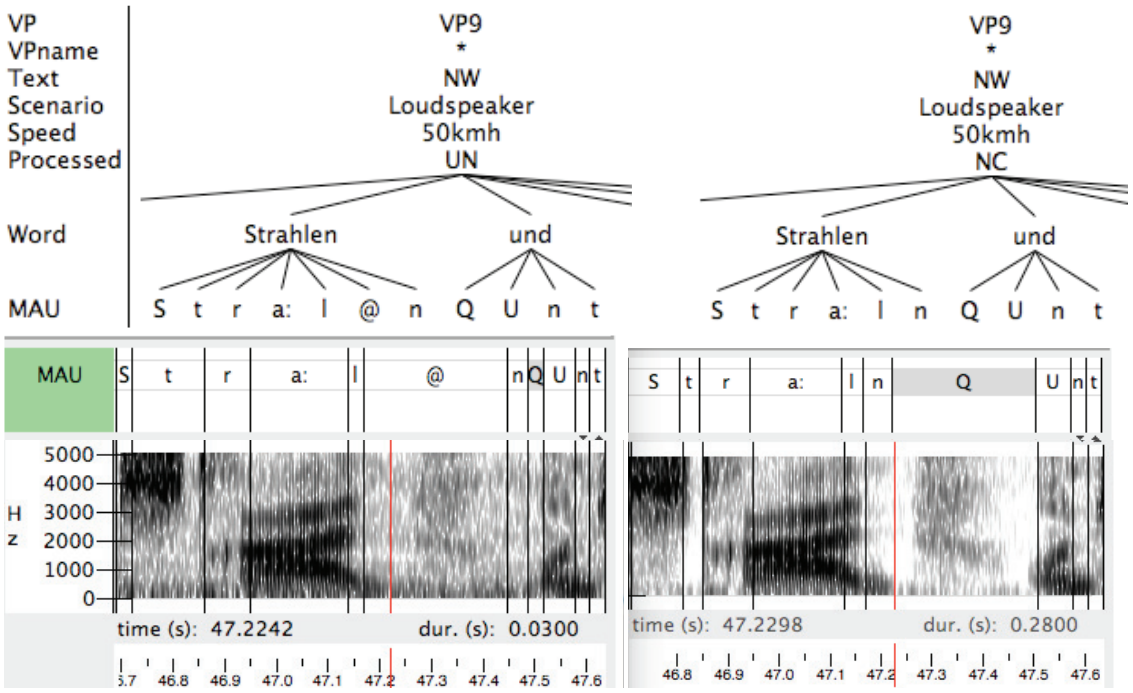

Figure 3 – Examples of the MAUS segmentation and labelling (MAU tier) aligned to the time signal (oscillogram and sonagram) together with the hierarchical annotation (trees) in EMU for a short part of a recorded unprocessed signal (left) and the same signal enhanced by NS (right)

All MAUS segmentations (for an example see Figure 3) from run 4 were compared across the different processing steps with the dependent variable recording situation (IC1-3 and REF) in terms of the beginning of the word segmentation (B_W), deviations in the chosen pronunciation

variants of words (V_W) and phones (V_P), as well as the duration and overall number of chosen phones (n_P, Δt_P) and pauses ($n_{<P:>}$, $\Delta t_{<P:>}$). Interferential statistics were used to interpret the results.

3 Results

Figure 3 exemplifies the outcome of the MAUS segmentation and labelling of the phones as well as the hierarchical labelling structure for words and some meta-data. In run 1 and 2 for which no prosodic information was given, MAUS segmentation failed (algorithm crashed or returned an error message) for most of the NC+S signals of IC3 and for some of the other speed simulations, as well as for some NC signals. In contrast, in run 3 (additional major prosodic units) MAUS segmentation succeeded for all signals but three (VP1 NC in IC3, VP2 NC+S in IC1, VP4 NC+S in IC2) and in run 4 (with minor prosodic units) for all but two (VP4 UN, VP6 NC both in REF). All UN signals were successfully segmented in all runs.

Figure 3 provides an example, how the selection of word pronunciation variants (V_W) and phones (V_P) can differ for the same utterance without and with signal enhancement (NC). While a [@n]-sequence was segmented in the UN signal in the word <Strahlen>, MAUS chose a word pronunciation with @-reduction in the NC signal. Furthermore, there are differences in phone lengths for the initial consonant cluster in the same word.

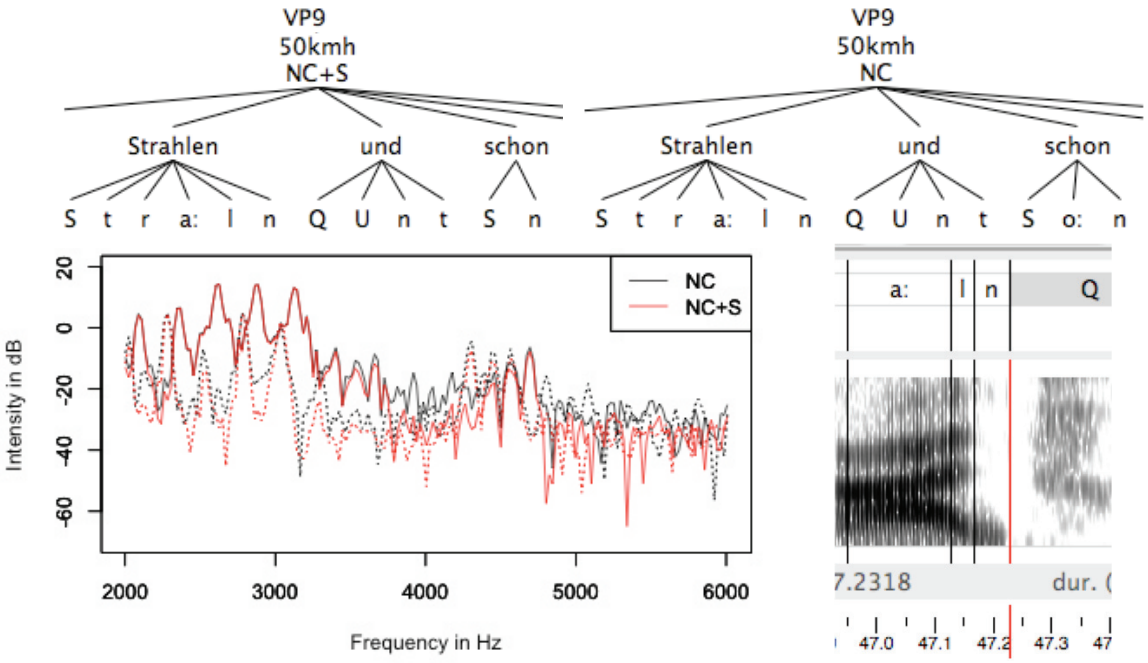

Figure 4 – Different word variants for NC and NC+S enhanced signals (top) and overlaid spectral slices (bottom-left) from the temporal midpoint of the phone following the postalveolar fricative (solid) and the nasal (dotted) in <schon> as well as the spectrogram of the nasal in <Strahlen> in the NC+S signal (bottom-right).

These kinds of differences are very small, i.e. they concern at most 12 deviations for the chosen word variants (between UN and NC signals over all segmentations of the recordings from the 100 km/h scenario). Deviations are also found in the recordings of the reference scenario (four times between UN-NC; six times between UN-NC+S). In all cases word variant differences are based mainly on unequal segmentations and labelling of glottal stops, alveolar stops, a-Schwa, e-Schwa and pauses. Over all, the simulation scenarios yielded more word variants with e-Schwa and less variants with a-Schwa, glottal stops and pauses for the UN than for NC and NC+S signals. The opposite was true for the reference scenario.

Figure 4 exemplifies differences in the segmentation in dependency of the different processing steps NC and NC+S. Another vowel than a-Schwa or e-Schwa is affected. The

spectral slices in Figure 4 considerably deviate between 4000-7000 Hz. This frequency range, which is crucial for nasal detection, is suppressed by NC+S to the extent that the intensity of the vowel section does not differ anymore from that of the following nasal (solid grey lines are in the area of the dotted lines), while the two sections do still differ at about 3800 Hz. Figure 3 also shows, how the formant structure of the nasal in <Strahlen> is maintained after NC processing, rather than being suppressed after NC+S processing, as in Figure 4.

The analysis of the beginning of the word segmentation (B_W) showed significant ($p < 0.001$) differences between the three steps of processing (UN, NC, NC+S). NC or NC+S processing did not affect B_W of the reference scenario recordings. However, for IC1-3 B_W was later in the NC and NC+S signals (differences NC-NC+S post-hoc $p=0.857$) than for the UN signals (differences to NC and NC+S each $p < 0.001$). In the UN condition, segmentation and labelling started quite at the beginning of the recording except for the 50 km/h scenario. From a descriptive point of view only, NC+S led to earlier B_W values than NC except for IC3.

Figure 5 exemplifies the differences in phone length (Δt_P) for the different processing steps. First of all, the whisker-box plots display quite equal mean values for all sound classes. The same is true for pauses (not shown in the figure). The differences between the steps reside in the outliers. For most of the classes, outliers are reduced in the enhanced signals, whereby NC reduces outliers to an extent comparable to that of the unenhanced reference signals. The exception is the vowel class. Here, outliers in terms of longer vowel durations occurred for both processing methods in contrast to the reference signal. The largest number of outliers was found in the unprocessed signals.

To sum it up, compared to the REF signals, B_W, V_P, n_P and Δt_P values are very different in UN signals of the IC1-3 conditions, whereas for the NC and NC+S signals, the values approximate those of the reference values in the REF signals.

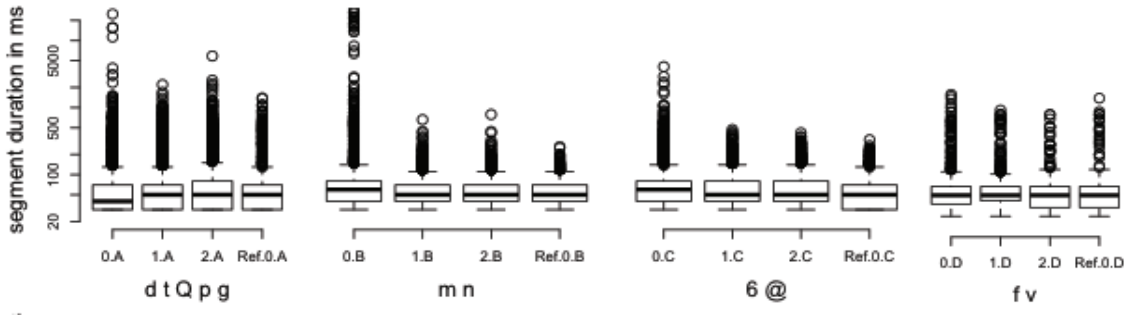

Figure 5 – Phone lengths in ms for different sound classes for all simulation scenarios together, separated by the 3 processing steps (UN=0, NC=1, NC+S=2) in comparison to the unprocessed reference scenario signals (Ref.0).

4 Discussion

The results show that MAUS segmentation succeeds only with a very detailed canonical transcription. The segmentation of the unprocessed signals in the first run has to be more erroneous than for the processed signals, because of the inaccurate canonical transcription. Manual corrections or precisions of the canonical transcriptions became necessary, since, for example, some of the speakers varied word forms or did not read the title. Unenhanced signals are inevitably impaired by noise. This noise led to misinterpretations in the MAUS' alignment procedure. Somehow surprising is the fact that the reference signals, which were not overlaid by noise, were segmented successfully in the first run. An explanation might be that the reference scenario was the first to be read by the speakers. The speakers did not know the narrative so that they read very slowly and articulated very clearly. More detailed linguistic analyses may shed some light on this question. Moreover, some speakers did not adhere to the punctuation when reading the narrative. These recordings make a further revision of the canonical transcription of prosodic units necessary.

The discussion above already implies differences in the word variants chosen by MAUS for the unprocessed signals on the one hand and the enhanced signals on the other. Yet, the additionally found differences between the two processing steps NC and NC+S are somewhat unexpected. Some of these differences may be due to the fact that the suppression method masks some phonetic cues, whereas the cancellation method does not. Other differences between the word variants selected by MAUS are explainable by the phonetic contexts in which the sound segments occurred. Glottal stops, a-Schwa and e-Schwa all appear at word boundaries and are thus often adjacent to pauses and/or affected by final lengthening and rapidly decreasing intensity. Sounds in such contexts are hard to detect in noisy environments and also hard to maintain by signal enhancement. Furthermore, e-Schwa deletion is a common phonological process in German, which is taken into account by MAUS. It seems plausible that MAUS did not segment an e-Schwa whenever there were no obvious phonetic cues to a vowel segment. The same is true for the alveolar stops. They appear in the most frequently occurring words in the narrative and are usually subject of articulatory reduction (e.g. <Wanderer> [vandəʁɐ] > [vanːəʁɐ]; <einst stritten> [ʔaɪnstʃtʁɪtən] > [a̰ɪnsʃtʁɪʔn̩]). So, MAUS is more likely not to segment and label these stops unless there are strong cues. But, vice versa, MAUS chooses the canonical word pronunciation variants with e-Schwa and alveolar stops in noisy parts, because noisy signals contain enough acoustic energy for the time period to align all labels in the graph.

The different word variants yielded by different enhancement steps lead to the conclusion that there are differences in the signal regarding the amount and clarity of acoustic cues that remain in the signal. These differences need to be taken into account in follow-up studies. With respect to the missing spectral structure (particularly in the mid-frequency region) in signals enhanced by NC+S, the NC enhancement method without suppression seems to be overall more appropriate.

The earlier beginning of the word segmentation for the simulation scenarios with the noisy signals shows that MAUS aligns phones to noisy parts of the signals even before the first phone is actually produced. Although there is no statistical confirmation of the differences between an earlier segmentation start in the NC+S than in the NC signals, the results show different effects on the MAUS performance in dependency of the enhancement method.

The choice of the pronunciation variant and the beginning of the word segmentation are closely related. MAUS already aligns phones to the noisy signal before the actual speech signal starts. The wrong word segmentation in the beginning of the utterance affects the word segmentation in the actual speech signal due to the graph based alignment method used by MAUS. The graph is not recycled, when parts are already aligned. Thus the wrong beginning causes follow-up errors up to an unpredictable time in the signal. Signal enhancement can correct these errors to a certain degree, primarily by avoiding an early segmentation start. The small but existing differences in NC and NC+S performances for the speed scenarios IC1-3 may be interpreted in the light of the acoustic energy left in the signal. The canonical transcription of the narrative started with pauses followed by a glottal stop. MAUS selected different onsets for the glottal stop. As has been explained in the results section, glottal stops detected by MAUS vary very much in duration. Thus, in our database these stops need to be treated as prosodic units rather than as phones. A more detailed canonical transcription may solve this problem. More detailed canonical transcriptions mean higher effort in the pre-processing. However, the benefits must be weighed against the higher effort.

Phone length is a further parameter, which heavily depends on the beginning of the word segmentation and the acoustic energy in the signal. The length of all the phones already aligned to the noisy part of the signal only show the characteristics of the graph used by MAUS, i.e. intervals of 10 ms. Acoustical energy is more strongly suppressed by noise suppression than by noise cancellation. Thus, differences in phone length are predictable as

far as phones adjacent to pauses are concerned. Due to the results on phone length noise cancellation seems to be more preferable.

5 Conclusion

The present results show that the time-consuming task of annotation can be automated for noisy signals when they are improved by noise cancellation and noise suppression methods. Further it shows that enhancement can substantially affect the phone length and the beginning of the word segmentation.

The results received after noise cancellation were better than those where additional suppression was applied to the noisy signals. The noise cancellation is a method that crucially depends on the knowledge about the acoustic transmissions from the loudspeakers to the microphones. Noise cancellation is thus not applicable in every recording situation, but limited to ambience simulation. However, in such a simulation our method is useful for gaining insights into spoken language communication in noisy environments, as in in-car communication. The present findings are a prerequisite for the next steps of our study, i.e. the phonetic and linguistic analyses of speech in noisy environments.

6 Outlook

Only a comparison of the MAUS segmentation and a manual segmentation is able to assess the precision of the set segment boundaries. However, manual segmentation was not carried out as yet. Further analyses need to confirm that the signal – despite its processing – contains enough acoustic cues to be comprehensible for listeners and already existing data extracting algorithms.

Further, additional recordings will contain spontaneous speech, including enhancement by noise suppression without preceding noise cancellation. The goal will be to enhance conventional noise suppression schemes such that they can also improve the results of phonetic signal analyses tools.

References

[1] Boersma, P. Praat, a system for doing phonetics by computer. Glot international 5, pp. 341–345, 2002.

[2] Jaschke, S. Generierung und Analyse einer Sprachdatenbank unter Berücksichtigung des Lombard-Effekts. Christian-Albrechts-Universität zu Kiel, 2012.

[3] John, T. EMU Speech Database System: Praxisorientierte Weiterentwicklung der Funktionalität, Benutzerfreundlichkeit und Interoperabilität sowie die Aufbereitung des Kiel Corpus als EMU-Sprachdatenbank. PhD Thesis, LMU Munich. 2012.

[4] John, T. and L, Bombien. EMU. In: Durand, J., Gut U. and G. Kristoffersen, editors, Handbook on Corpus Phonology. Oxford University Press, in press.

[5] Kipp, A., Wesenick, M. and F. Schiel. Automatic detection and segmentation of pronunciation variants in German speech corpora. In Spoken Language, 1996. ICSLP 96, volume 1, pp. 106–109.1996.

[6] Kipp, A., Wesenick, M. and F. Schiel. Pronunciation modelling applied to automatic segmentation of spontaneous speech. In: Proc. EUROSPEECH, 1997. pp. 1023–1026. 1997.