Time-domain Noise Subtraction Applied in the Analysis of Lombard Speech

Hansjörg Mixdorff, Katja Grauwinkel* and Martti Vainio***

*Department of Computer Sciences and Media, Berlin University of Applied Sciences, Germany
{mixdorff;grauwinkel}@tfh-berlin.de

**Department of Speech Sciences, University of Helsinki, Finland
martti.vainio@helsinki.fi

Abstract

This paper presents results of the comparison between speech produced in silence and speech in noise, also known as Lombard speech. A temporal filtering algorithm was developed which successfully removes the ambient noise from recordings of Lombard speech by locating and subtracting a recording of the noise performed in the same environment. The filtering algorithm yields overall noise attenuation between 15 and 30 dB without distorting the speech signal like spectral filtering approaches. In the subsequent acoustic analyses we examined the effect of varying levels of noise on vowel formants, glottal spectra and intensity. For most vowels we found significant rises in F1 and F2, but little variation in formant bandwidth. The overall rise in intensity between silent and 80 dB babble noise conditions was found to be of 9 dB. With growing effort higher harmonics are boosted by up to 6 dB whereas the average speech rate only drops by 5%. In Lombard speech the standard deviation of phone intensity is reduced.

1. Introduction

It is commonly known that humans in noisy environments adapt their manner of speaking. This adaptation not only concerns the loudness of speech, but also fundamental frequency, speech rate and spectral characteristics (for a summary see, for instance, [1]). The so-called 'Lombard' speech has been shown to be more intelligible than normal speech, and therefore raised the interest of researchers who aim at improving the quality of speech synthesizers. Furthermore, speech recognizers in a noisy environment have to deal with the properties of Lombard speech. Therefore suitable training corpora need to be developed. The current paper discusses a filtering approach designed to remove ambient noise from recordings of Lombard speech made in an anechoic chamber. Most current studies avoid the contamination of recordings of Lombard speech by presenting the noise to the subjects over headphones. Provided the headphones are closed type ones, the noise that the recording microphone will pick up is minimal. On the downside, this method requires that the talker's proprioception, that is, the sound of his own voice talking, and the acoustics of the room, are somehow simulated and also presented over the headphones. Besides, wearing closed type headphones for a longer period of time is tiresome and unnatural.

Commonly denoising is performed in the frequency domain by estimating the spectrum of the noise during speech pauses and subsequent spectral subtraction. This works fairly well for noise types with constant spectral properties while babble noise is more difficult to remove. Furthermore the speech signal is degraded by this kind of technique.

The great advantage of the lab condition is the fact that the properties of the contaminating noise are exactly known since it is presented to the listener from a recording. In a recent approach [2] the transfer function of the recording set-up is therefore estimated using white noise and a corresponding FIR filter is constructed. It requires that the talker be present in the room while the white calibration noise is recorded. On a two-channel recording the original of the contaminating noise and the noise contaminated speech recorded in the anechoic room are stored. Subsequently the channel of the contaminating noise is passed through the FIR filter and its output subtracted from the speech signal. This method yields typical attenuation levels of 30 dB. However, it requires sophisticated channel estimation techniques and specialized DSP hardware. In the current study we therefore examined a simpler method based on direct noise subtraction in the time domain.

2. Filtering Procedure

As in [2] we assume that the recording set-up is highly linear, time-invariant and low in noise, and therefore its transfer function remains fairly constant. A recording of the contaminating noise is performed while the talker is already present in the room. On subsequent takes, the talker starts speaking shortly after noise onset. After a set of recordings is completed, the initial noise-only recording (henceforth 'sample noise') is subtracted from the contaminated speech samples. To this effect, the sample noise must be precisely aligned with the speech recording. Therefore a search frame of 0.25 s is extracted shortly after the onset of the noise sample and searched for in the noise-contaminated speech file by determining the maximum cross-correlation between the search frame and subsequent frames of equal size from the speech file. Figure 1 shows an example of a cross-correlation function (CCF). The local maximum where the onset of the search frame was located in the speech file is clearly visible on the left hand side. In the right-hand third of the figure we see the onset of the speech signal marked by an overall increase in the CCF. This occasionally leads to wrong peak-picking and subsequent failure of the filtering method when the onset of the speech signal is located very closely after the beginning of the contaminating noise.

The filtering procedure was implemented as a MATLAB program and yields typical attenuation levels of 20 and up to 35 dB. We found that at a sampling rate of 48 kHz an alignment mismatch by only one sample to the left or right of the cross-correlation maximum causes a reduction of maximum attenuation by up to 10 dB. In theory the maximum

alignment mismatch amounts to half a sampling period. In a series of preliminary tests we found a variation of maximum attenuation of about +/- 2 dB. More detailed figures will be presented in the following section.

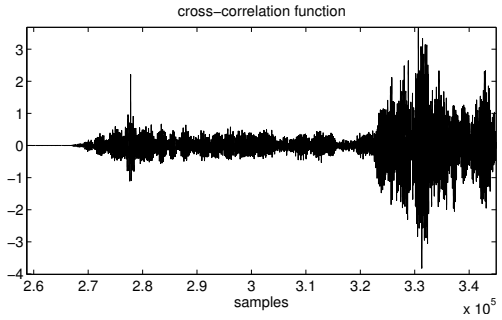

Figure 1*: Example of cross-correlation function (CCF) used for locating the babble noise in the recording. The local maximum on the left hand side marks the onset of the search frame in the recording.*

3. Speech Material

The speech material used in this study consisted of phonetically balanced German [3] and Finnish sentences between six and twelve syllables long. These were produced by one native speaker of German and one of Finnish, respectively, in the sound-proof recording room at the Department of Speech Sciences, University of Helsinki at 48 kHz/24 bit using an AKG C4000 condenser microphone connected to a Macintosh G5 via a Digidesign 002 Firewire interface. The microphone was located 20 cm from the talker's mouth. We tested four conditions: no background noise (henceforth referred to as 'normal'), and babble noise at 60, 70, and 80 dB SPL at the location of the recording microphone, and performed three repetitions. We adjusted the recording level to accommodate the 80 dB condition and left it unchanged for all remaining ones. Every recording subset consisted of sixteen sentences. In conditions 60, 70 and 80 dB babble noise was played back over loudspeakers of type Genelec 1029A with all tone controls off placed in front of the talker. The speakers were situated about 150 cm above ground and 40 cm off the wall about 100 cm apart. The distance to the talker's mouth was about 120 cm. In order to avoid adaptation the duration of the noise was limited to five seconds. At the first presentation in a set the noise signal was recorded alone, yielding the sample noise. Subsequently, with every following presentation of the noise the talker uttered one of the sentences. There was a pause of approximately five seconds between consecutive noise presentations.

Figure 2 displays a speech sample in condition 80 dB before (top) and after (bottom) filtering, the noise attenuation in this case is 29.9 dB. In the top panel the five seconds of babble noise contaminating the utterance produced by the talker are clearly visible. As we used a cardioid characteristic with a typical rear attenuation of 20 dB and the loudspeakers were mounted in front of the talker the amplitude of the babble noise is relatively low at an SNR of approx. 15 dB.

We observed that the noise attenuation depended on the noise condition and the position of an utterance in a set. For 60, 70 and 80 dB we yielded an utterance-wise mean attenuation of 17.4, 21.2 and 21.6 dB, respectively. In the case of 80 dB, for instance, the attenuation dropped from initial values of over 30 dB for the first utterances in a set to barely 20 dB for the last ones. These figures compare to those presented in [2]. Here it was observed that minimal displacements of the talker and even breathing caused changes in the transfer function of the sound-proof booth and therefore reductions in the maximum attenuation as the recording session progressed. It is more difficult to explain why the mean attenuation in the 60 dB condition is considerably lower than in the 70 and 80 dB conditions. We assume, however, that as the energy of the contaminating noise becomes smaller, random factors such as quantization noise and thermal noise which differ between the sample noise and the noise contaminated speech recording set limits to the noise subtraction method.

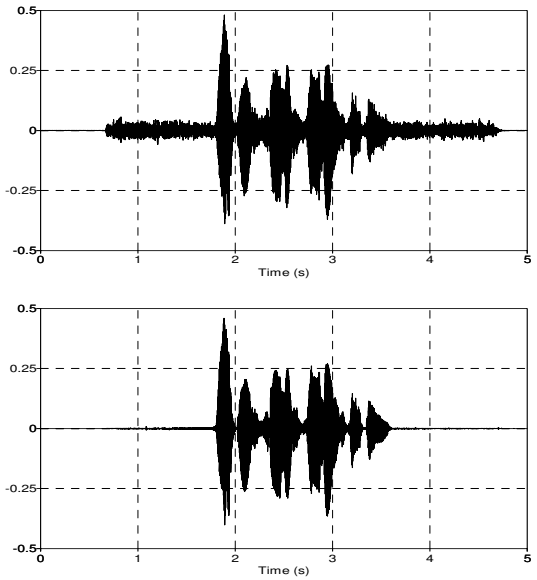

Figure 2*: Noise-contaminated speech sample before and after filtering, condition 80 dB. In order to avoid listener adaptation, babble noise was presented for five seconds only.*

In all cases examined, however, the noise reduction was sufficiently effective to carry out subsequent acoustic analyses, and the contaminating noise became practically inaudible. Furthermore, the speech signal yielded was clear and unaffected by the filtering procedure.

4. Acoustic Analysis

After filtering, speech samples were down-sampled to 16 kHz, annotated by forced alignment on the phone level and labels manually corrected. We determined formant frequencies and bandwidths at vowel centers using *Wavesurfer* [5], intensity contours as well as fundamental frequency contours at a step of 10 ms using *Praat* [6] default settings. The data was inspected and if necessary corrected.

Table 1 lists mean F1 and F2 values for a set of German monophthong vowels. As can be seen, F1 increases

significantly as the noise level rises. F2 increases slightly for most of the vowels, but the picture is not consistent.

Table 1: *Mean formant frequencies calculated for monophthong German vowels.*

Phone	F1/F2 [Hz] normal	F1/F2 [Hz] 60 dB	F1/F2 [Hz] 70 dB	F1/F2 [Hz] 80 dB
[2:]	385/1511	406/1440	412/1480	444/1353
[9]	424/1398	427/1432	427/1581	465/1602
[a]	571/1286	597/1304	627/1325	674/1370
[a:]	664/1187	675/1197	727/1224	776/1266
[e:]	398/1854	405/1908	411/1963	454/1978
[E]	477/1695	481/1707	514/1714	556/1759
[i:]	318/2045	327/2043	336/2055	332/2096
[I]	369/1716	376/1674	406/1784	443/1772
[o:]	397/1084	404/1073	420/1127	471/1225
[O]	518/1034	545/1013	570/1072	609/1070
[u:]	331/943	345/959	364/1036	407/1120
[U]	442/1046	440/1075	486/1113	510/1023
[y:]	320/1679	339/1633	366/1575	332/1577
[Y]	390/1445	418/1453	439/1449	450/1435

The formant bandwidth measurements yield differences between the four different conditions as can be seen in Table 2, but the mean bandwidths do not show any consistent tendency with respect to the noise condition. In particular we do not see any consistent reduction in bandwidth under Lombard conditions as reported in [7], for instance. Similar results can be seen in Table 3 for the Finnish vowels. Figure 3 shows how the vowel plane is shifted under Lombard conditions, especially in the case of the German speaker.

Table 2*: Mean formant bandwidths calculated for a number of monophthong German vowels.*

Phone	B1/B2 [Hz] normal	B1/B2 [Hz] 60 dB	B1/B2 [Hz] 70 dB	B1/B2 [Hz] 80 dB
[a:]	185/98	160/85	164/78	270/119
[e:]	76/156	86/164	94/136	123/139
[o:]	74/132	66/164	72/106	57/105
[u:]	87/174	83/153	93/165	92/144

Analysis of intensity contours yields overall rises by 1.45, 4.85 and 8.85 dB for conditions 60, 70 and 80 dB, respectively, as compared to normal speech. If we only consider the intensity in the middle of each segment, the figures are 2.51, 6.19, and 10.24 dB, respectively. This rise concerns vowels (2.25, 6.03, 10.41 dB) and consonants (2.82, 6.07, 10.01 dB) almost equally. As the mean intensity rises, the standard deviation becomes smaller, from 5.37 dB in the case of vowels in normal condition to 3.56 dB in condition 80 dB. We also observe that the usual intensity difference between lax and tense vowel types ([2:]/[9], [E]/[e]:, [I]/[i:], [O]/[o:], [U]/[u:], [Y]/[y:]) is leveled.

In order to determine how the spectrum of the glottal source is affected by the Lombard condition we inversely filtered all vowel segments (LPC of order 16) and pitch-normalized the resulting residual signal to 160 Hz using PSOLA techniques within *Praat*. The resulting averaged spectra for the four conditions can be seen in Figure 4. As can be seen, as overall vocal effort increases from the normal condition to 80 dB higher harmonics are boosted up to 5 dB compared to the fundamental.

Table 3*: Mean formant frequencies calculated for monophthong Finnish vowels.*

Phone	F1/F2 [Hz] normal	F1/F2 [Hz] 60 dB	F1/F2 [Hz] 70 dB	F1/F2 [Hz] 80 dB
[a]	640/1306	652/1298	679/1296	699/1323
[ä]	647/1533	651/1516	660/1533	675/1564
[e]	459/1727	490/1718	497/1738	513/1794
[i]	365/1901	391/1850	392/1877	403/1918
[o]	473/1107	484/1100	500/1130	524/1143
[ö]	499/1449	493/1443	493/1477	530/1506
[u]	355/859	371/848	388/920	399/897
[y]	330/1755	353/1702	376/1713	381/1743

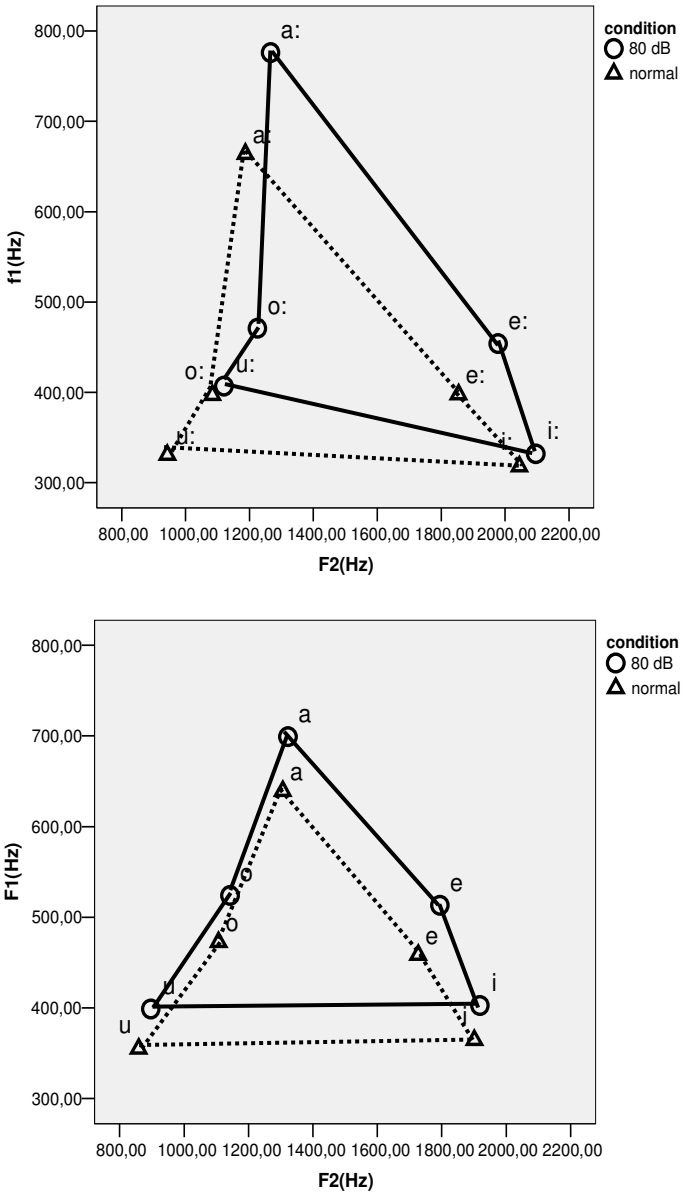

Figure 3: *Formant chart of German (top) and Finnish (bottom) vowels.*

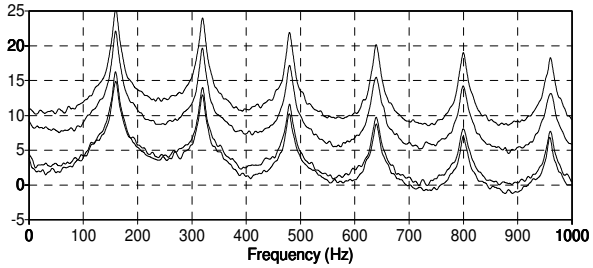

Figure 4: *Spectra of pitch-normalized LP-residuals for normal speech and babble speech and babble noise at 60, 70 and 80 dB (from bottom to top).*

The fundamental frequency contours produced by the German speaker were parameterized using the Fujisaki model [8] maximally assigning each accented syllable one accent command as outlined in [9]. Figure 5 displays two examples of analysis: Utterances of the sentence "Am blauen Himmel ziehen die Wolken." - "The clouds are drifting in the blue sky." are shown in conditions normal (top) and 80 dB noise (bottom). As can be seen, Lombard speech is characterized by increased accent command amplitudes *Aa* and also increased phrase command amplitudes *Ap*. Furthermore, the last accented syllable [vOl] carries a high tone in 80 dB as compared to a low tone in the normal condition requiring an additional accent command. Table 4 displays means and standard deviations for *Aa* and *Ap* for all four conditions. Generally speaking, rising noise levels are accompanied by stronger resets of the declination line as captured by *Ap* and higher accent prominence as captured by *Aa*. The total number of accent commands assigned rises from 170 under normal condition to 200 in 80 dB, mostly because of the sentence-final accents becoming associated with high tones.

Table 4: *Means and standard deviations of accent command amplitude Aa and phrase command magnitude Ap for all four conditions.*

	mean/s.d. normal	mean/s.d. 60 dB	mean/s.d. 70 dB	mean/s.d. 80 dB
Aa	.34/.15	.38/.15	.39/.15	.45/.17
Ap	.44/.13	.45/.12	.57/.10	.69/.16

5. Discussion and Conclusions

The present study introduced a simple filtering method for noise-contaminated Lombard speech. Its main advantage is that talkers do not have to wear headphones during recording, impeding their proprioception. By pre-recording and subsequently locating and subtracting the noise, attenuation levels of up to 30 dB can be yielded. The method was tested on recordings of phonetically balanced sentences by one German and one Finnish speaker and under conditions of babble noise values of 60, 70 and 80 dB. The filtered speech samples were clear enough to perform acoustic analyses such as formant estimation and pitch extraction. Analysis results were in line with earlier studies of Lombard speech regarding f0, F1, F2 and intensity increases [1], but we could not observe a consistent reduction of formant bandwidth as reported in [7]. This result might be explained by the higher noise levels of 95 dB SPL used which causes subjects to actually shout. In future works we would like to examine the effect of adverse conditions on communicative strategies in a task-oriented dialog.

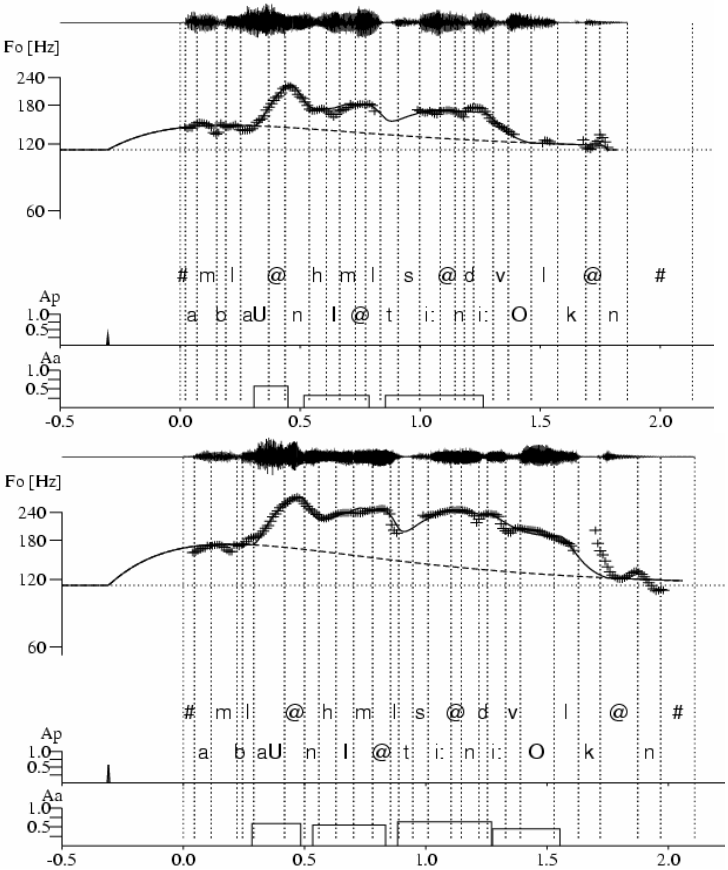

Figure 5: *Examples of Fujisaki model-based F0 contour analysis for normal (top) and Lombard speech (bottom).*

6. References

[1] Junqua, J.-P. 1996. The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. *Speech Communication* 20, 13-22.

[2] Ternström, S. Södersten, M.; Bohman, M., 2002. Cancellation of simulated environmental noise as a tool for measuring vocal performance during noise exposure. *Journal of Voice* 16(2), 195-206.

[3] Sotschek, J. 1984. Sätze für Sprachgütemessungen und ihre phonologische Anpassung an die Deutsche Sprache. Tagungsband DAGA: Fortschritte der Akustik, pp. 873-876.

[4] Vainio, M. et al. 2005. Developing a speech intelligibility test based on measuring speech reception thresholds in noise for English and Finnish, JASA2005, 118(3),1742-1750

[5] www.speech.kth.se/wavesurfer

[6] www.praat.org

[7] Bořil, H. and Pollák, P., 2005. Design and Collection of Czechk Lombard Speech Database. In *Proceedings of ESSV2005*. Prague.

[8] Fujisaki, H., Hirose, K., 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. J. of the Acoustical Society of Japan (E), 5 (4), 233-241.

[9] Mixdorff, H. 2001. MFGI, a Linguistically Motivated Quantitative Model of German Prosody. In Improvements in Speech Synthesis, E. Keller et al. (Ed.), Wiley Publishers, pages 134-143, UK.