ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

REMOVING HUM FROM SPOKEN LANGUAGE RESOURCES

Hartmut R. Pfitzinger

Department of Phonetics and Speech Communication
University of Munich, Schellingstr. 3, 80799 München, Germany
hpt@phonetik.uni-muenchen.de

ABSTRACT

This paper presents a technique to remove the hum in speech corpora based on time domain subtraction rather than spectral subtraction. This novel method is able to avoid any perceptible or measurable artifacts and has proved to be an efficient technique for completely eliminating even complex hum waveforms from speech recordings.

1 INTRODUCTION

In recent years, two widespread concepts of noise reduction algorithms could be observed: spectral noise subtraction and adaptive filtering. The former has the drawback of generating residual noise with musical character, the so-called musical noise [1, 9, 10], while the latter distorts the frequency and phase response of speech signals [4, 6, 7]. In addition, both methods fail to enhance speech recordings disturbed by loud hum (50 Hz or 60 Hz) because their analysis windows enlarge the line spectrum by significant artificial side lobes.

A serious problem emerges when hum is clipped, which often arises from recording equipment with poor shielding and/or grounding. This causes complex hum waveforms with odd harmonics that occasionally match the fundamental frequency of the speaker and some of its harmonics. In contrast to a single 50 Hz or 60 Hz sinusoidal hum wave, which could be removed by high-pass or Notch filtering, comb filtering of complex hum waves leads to unacceptable distortions of the transfer function of the speech signal.

2 SYSTEM OVERVIEW

To overcome these problems we developed a new subtraction method for removing any stationary harmonic noise, particularly hum which is additive, independent of the speech signal, and is an instance of this class of disturbances. The method is based on temporal rather than spectral subtraction, hence the signal of the hum is needed in isolation. This we obtain by using additive synthesis with restricted frequency ratios, where each sinusoidal component is frequency-, amplitude-, and phase-locked to each detectable harmonic of the hum.

Detection and measurement of the harmonics of the hum is done by the constraint-analysis module of our algorithm which is specially designed to search the input signal for a stationary acoustic object with a fundamental frequency of 45 Hz to 65 Hz and with odd harmonics. It analyses the non-speech signal passages assuming that they are sufficiently marked by low spectral energy in the frequency range of 400 Hz to 3.6 kHz. The analysis step provides detailed information about whether there is hum or not, about the hum-to-noise ratio, and a list of all hum harmonics with frequency, amplitude, and phase information.

An additive synthesis module uses the information provided by the constraint-analysis module to create a copy of the hum from the speech signal. Subtracting the re-synthesized hum signal from the speech signal results in a highly enhanced version of the speech signal because hum is a purely additive disturbance.

3 NON-SPEECH DETECTION

For detecting non-speech passages in signals a FIR bandpass filter is designed which reduces the amplitudes of the first five harmonics of a 50 Hz hum by at least 40 dB [2, 5, 8]. Fig. 1 shows its frequency response. The amount of passband ripple of actually ± 0.4 dB is not decisive, while the low cutoff edge should be as sharp as possible to minimally reduce low-frequency energy of speech.

A bandpass filter instead of a lowpass filter is used to prevent low- as well as high-frequency disturbances from distorting the speech/non-speech decision. In addition, the hum removal system should ignore frequencies above 3.6 kHz allowing for its application to telephone recordings and to speech corpora sampled at only 8 kHz. Convolution with the FIR bandpass filter enables detection of speech signal passages, since speech shows significant spectral energy above the low cutoff frequency of the filter (400 Hz) while hum typically has little spectral energy at the 8th or higher harmonics.

The effect of the FIR bandpass filter on the short-term amplitude estimation is shown in Fig. 2: the filtered speech signal lost some energy because low harmonics of speech are affected by the transfer function. But the dynamic range is significantly improved, e.g. allowing the distinction between silence and the laterally released initial stop of the utterance [kˡaɪnɐ].

The enhancement of the dynamic range is also shown in the histograms of Fig. 3. They are derived from 5202 utterances spoken by 51 subjects. While the original recordings only have a short-

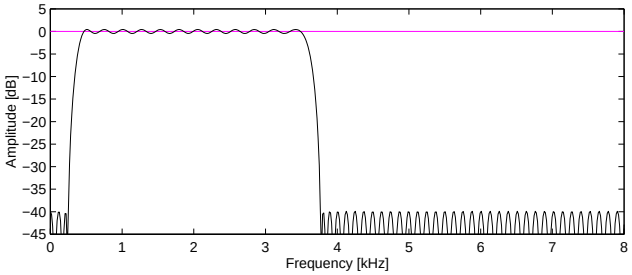

Figure 1: Transfer function of the FIR bandpass filter.

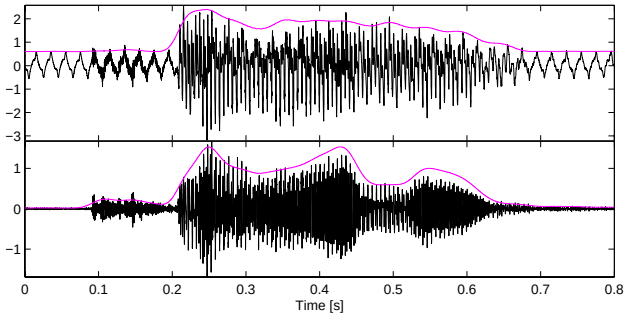

Figure 2: Original speech signal (top) and FIR bandpass filtered speech signal (bottom). Light lines show the short-term amplitude. Utterance: [kˡaɪnɐ] *kleiner* 'smaller'.

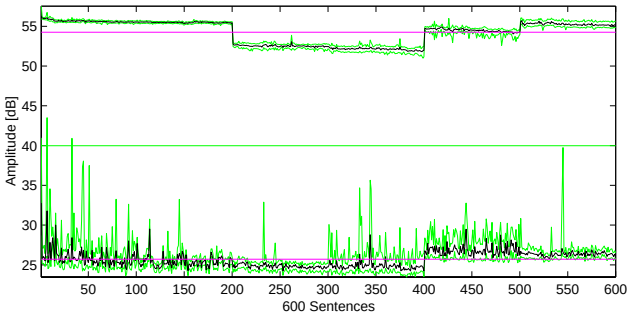

Figure 4: 25th percentiles (black) and 5th to 50th percentiles (light) of short-term amplitude distributions of 600 original (upper) and FIR bandpass filtered (lower) speech signals.

term amplitude range of less than 20 dB, the FIR bandpass filtered signals achieve a dynamic range of nearly 45 dB. In Fig. 4 a detailed analysis of the hum level of 600 out of 5202 utterances is shown which suggests a threshold of about 40 dB to distinguish between speech and non-speech signal passages.

4 HUM ANALYSIS

Applying a threshold of 40 dB to the short-term amplitudes of FIR bandpass filtered speech signals allows for accumulating speech and non-speech signal passages. In Fig. 5 the mean spectral energy distributions of these two signal types are shown.

Our hum removal system needs only two non-speech passages each with a duration of at least 100 ms to estimate the fundamental frequency of the hum. Since the distance between these two passages should be as large as possible the threshold search starts from both the beginning and the end of a signal. If the search algorithm detects only one or no non-speech passage, the user could redefine the threshold and minimum duration criteria. The search yields an index function $n = \mathrm{Ix}(m)$ where $m = 1, 2, \ldots, M$ results in $n \in \{0, 1, \ldots, N-1\}$ which are indices to samples of non-speech passages of the speech signal $S(n)$ with $n = 0, 1, \ldots, N-1$. Note that $m_1 > m_2 \Rightarrow \mathrm{Ix}(m_1) > \mathrm{Ix}(m_2)$. The non-speech signal to be analysed is obtained by $s(\mathrm{Ix}(m)) = 0$ and $s(j) = S(j)$ where $j \in \{0, 1, \ldots, N-1\}$ and $j \neq \mathrm{Ix}(m)$.

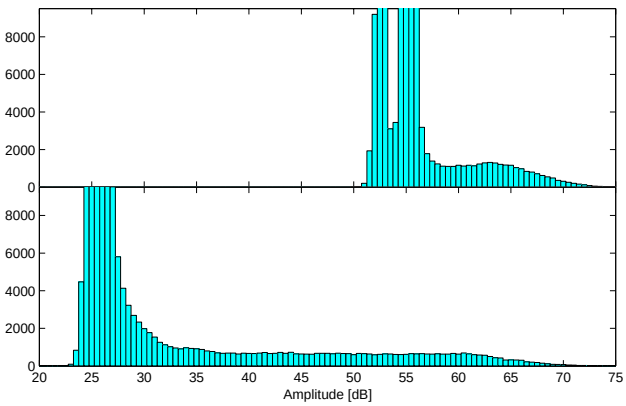

Figure 3: Histogram of short-term amplitudes of original speech signals (top) and FIR bandpass filtered speech signals (bottom).

Passages of the speech signal which are not in $\mathrm{Ix}(m)$ are not necessarily speech passages. Apart from speech they could contain transient disturbances and other non-speech passages in the middle of the signal. Thus, in Fig. 5 the mean spectral energy distribution of speech is a bit contaminated and consequently partially underestimated in contrast to the mean spectral energy distribution of non-speech which is valid.

Three methods for analysing frequencies, phases, and amplitudes of the hum harmonics were informally compared: (i) peak-picking in the spectral domain and an optimization algorithm based on golden section search and parabolic interpolation [3] were used to estimate each harmonic independently (Fig. 6), (ii) first (i) was applied and then each recognized harmonic was removed from the signal before analysing the next harmonic, (iii) only odd harmonics $k = 1, 3, ..., K$ were jointly analysed introducing a constraint to only use integer ratios between fundamental frequency and harmonics:

If $s(n)$ with $n = 0, 1, \ldots, N-1$ is the discrete-time non-speech signal to be analysed consisting of N samples, then we can estimate the cumulative amplitude A_f of hum and its harmonics with a fundamental frequency f by

$$A_f = 20 \sum_{l=0}^{\frac{1}{2}(K-1)} \log \left| \frac{1}{N} \sum_{n=0}^{N-1} s(n)\, e^{i\pi(2l+1)n\frac{f}{f_S}} \right|$$

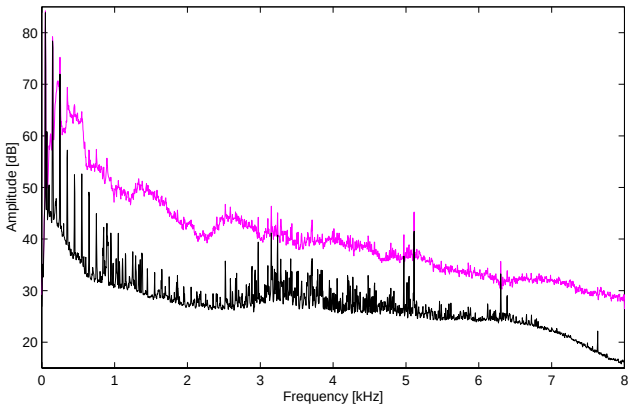

Figure 5: Mean spectral energy distribution of speech (upper) and of non-speech signal passages (lower) estimated from 600 original utterances. Notice the harmonics of the hum.

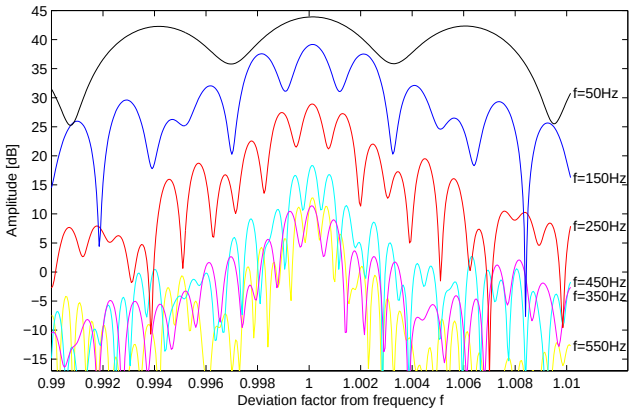

Figure 6: Spectral analysis of non-speech signal passages of an original utterance at frequency ranges 50 Hz ±1%, 150 Hz ±1%, ..., 550 Hz ±1%. Notice that the energy peaks of all harmonics of the hum match at multiples of 50 Hz.

where K is the highest odd harmonic to be analysed and f_S is the sampling frequency. A typical result of this analysis method for $f = 45 \ldots 65$ Hz is shown in Fig. 7. The precise fundamental hum frequency is found by using this formula together with an optimization algorithm based on golden section search and parabolic interpolation [3].

Method (iii) with the constraint of using integers as the only allowed multiples between the fundamental frequency and the harmonics was best to analyse the hum signal. The reason is that the harmonics (i.e. the odd harmonics) are a result of electrical distortion and therefore show exact integer ratios in contrast to rational spread factors between harmonics which emerge from, for example, physical properties of acoustic musical instruments, e.g. a piano.

Since in methods (i) and (ii) the frequencies of the harmonics are determined by searching for spectral peaks they could deviate by small amounts from the real harmonic frequencies. Hence the modelled hum spectrum could not match the real hum spectrum perfectly which is essential for removing hum without introducing any artifacts. Fig. 6 shows clearly that a frequency analysis error of 0.1% from the real fundamental hum frequency could lead to an amplitude analysis error of up to 6 dB.

A sample result of amplitude, frequency, and phase information derived by our hum analysis method is shown in Fig. 8. The main output of the method is the fundamental hum frequency f_H.

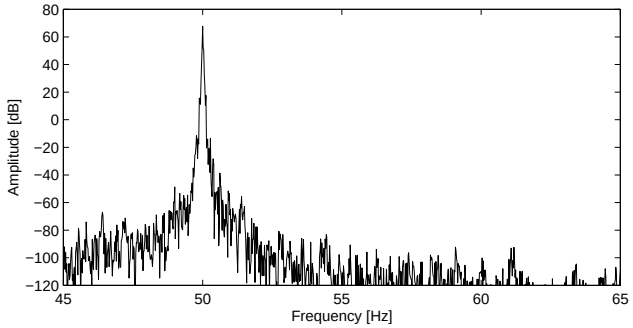

Figure 7: Result of the new constraint hum analysis method.

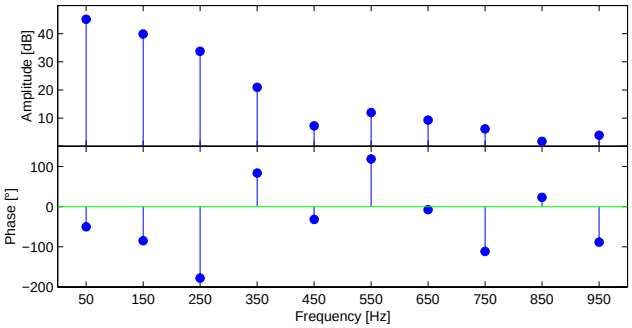

Figure 8: Sample output of the hum analysis algorithm.

5 HUM SYNTHESIS

If $s(n)$, $n = 0, 1, \ldots, N-1$ is the discrete-time non-speech signal to be analysed consisting of N samples, then we can synthesize sinusoid signals $y_k(n)$, which are frequency-, amplitude- and phase-locked to the k-th harmonic of the hum frequency f_H in the signal $s(n)$ by the following equations:

$$x_k(n) = e^{i\pi k n \frac{f_H}{f_S}}$$

$$a_k = \frac{1}{M} \sum_{n=0}^{N-1} x_k(n)\, s(n)$$

$$ci = \frac{1}{M} \sum_{m=1}^{M} \mathrm{Im}(x_k(\mathrm{Ix}(m)))^2, \quad cr = \frac{1}{M} \sum_{m=1}^{M} \mathrm{Re}(x_k(\mathrm{Ix}(m)))^2$$

$$y_k(n) = \frac{\mathrm{Im}(a_k)}{ci} \mathrm{Im}(x_k(n)) + \frac{\mathrm{Re}(a_k)}{cr} \mathrm{Re}(x_k(n))$$

where $n = \mathrm{Ix}(m)$ with $m = 1, 2, \ldots, M$ are the M indices of non-speech samples, and f_S is the sampling frequency. The correction factors ci and cr are important since N is not necessarily a multiple of the period duration of $x_k(n)$.

Finally, subtracting the synthesized sinusoid signals $y_k(n)$ from the original speech signal $S(n)$ leads to a perfectly enhanced speech signal $S'(n)$ (for example compare Fig. 11 with Fig. 2).

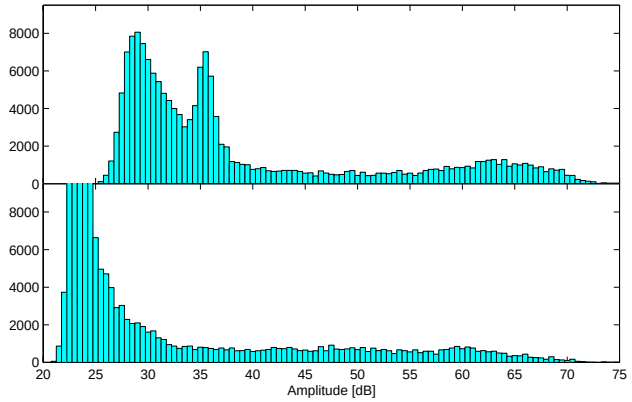

Figure 9: Histogram of short-term amplitudes of enhanced speech signals (top) and FIR bandpass filtered enhanced speech signals (bottom). Compare with Fig. 3.

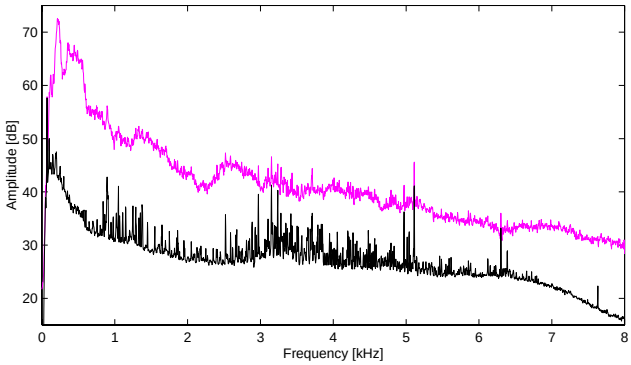

Figure 10: Mean spectral energy distribution of speech (upper) and of non-speech signal passages (lower) estimated from 600 enhanced utterances. Compare with Fig. 3.

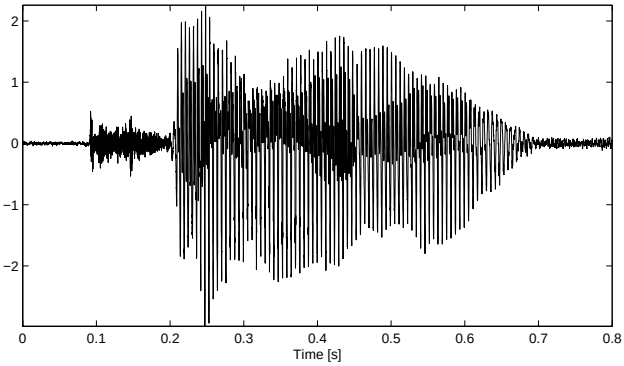

Figure 11: Enhanced speech signal. Utterance: [kl̥aɪnɐ] *kleiner* 'smaller'. Compare with Fig. 2.

6 EVALUATION

The spoken language corpus used for evaluation was recorded with damaged recording equipment in an office environment. The corpus consists of 102 utterances, each spoken by 51 speakers, resulting in 5202 speech recordings with a mean duration of 5 seconds. We applied our system to the whole corpus and then randomly selected 600 enhanced sentences for comparison with the original recordings.

Fig. 10 in comparison to Fig. 5 shows clearly that the processed harmonics in the 0–1 kHz frequency range are completely eliminated. The spectral peaks around 1 kHz reflect additional disturbances which have nothing to do with hum. The same applies to the single spectral peak at ca. 55 Hz. These peaks result from some of the 600 sentences which represent specific sound characteristics of the office environment.

The histograms (Fig. 9 vs. Fig. 3) indicate that signal-to-hum ratio could be improved by at least 20 dB comparing the peaks in the histograms of unfiltered signals. Even in the case of the FIR bandpass filtered speech signals (bottom histograms) a small improvement of 3 dB could be observed. Moreover, in the time domain the enhancement is obvious (Fig. 11 vs. Fig. 2).

Neither informal perception tests nor additional temporal or spectral analysis with various analysis methods (F0-detection, FFT, ARMA, and our constraint-analysis) were able to discover any residual hum in enhanced speech signals or even to uncover signal processing artifacts introduced by this new method.

CONCLUSION

The new method of subtracting an exact re-synthesized time domain copy of a complex hum signal from an originally hum-disturbed speech signal was able to completely remove the hum from speech signals. The hum was reduced to unmeasurable and imperceptible levels without introducing any artifacts. This method has proven to be extremely effective and valuable for spoken language corpus production, restoration, and maintenance.

Future improvements of this method should concentrate on non-stationary hum and other harmonic disturbances. In particular, the recognition of the class of disturbance is currently still left to the user to perform.

REFERENCES

[1] Cappe, O. (1994). Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Transactions on Speech and Audio Processing*, 2(2): 345–349.

[2] Digital Signal Processing Committee of IEEE Acoustics, Speech and Signal Processing Society, ed. (1979). *Programs for digital signal processing*. IEEE Press, New York.

[3] Forsythe, G. E.; Malcolm, M. A.; Moler, C. B. (1976). *Computer Methods for Mathematical Computations*. Prentice-Hall, Englewood Cliffs; New Jersey.

[4] Kenny, O. P.; Nelson, D. J. (1998). A model for speech reverberation and intelligibility restoring filters. In *Proc. of ICSLP '98*, vol. 2, pp. 479–482, Sydney; Australia.

[5] Madisetti, V. K.; Williams, D. B., eds. (1998). *The Digital Signal Processing Handbook*. CRC Press (IEEE Press), Boca Raton; Florida.

[6] Pfitzinger, H. R. (1997). Die Berechnung digitaler FIR-Entzerrfilter aus Messungen analoger Filter. Forschungsberichte (FIPKM) 35, pp. 185–190, Institut für Phonetik und Sprachliche Kommunikation der Universität München.

[7] Pfitzinger, H. R. (1998). The collection of spoken language resources in car environments. In *ICLRE '98*, vol. 2, pp. 1097–1100, Granada; Spain.

[8] Rabiner, L. R.; Gold, B. (1975). *Theory and application of digital signal processing*. Prentice-Hall, Englewood Cliffs; New Jersey.

[9] Yoma, N. B.; McInnes, F. R.; Jack, M. A. (1997). Spectral subtraction and mean normalization in the context of weighted matching algorithms. In *Proc. of EUROSPEECH '97*, vol. 3, pp. 1411–1414, Rhodes; Greece.

[10] Yoma, N. B.; McInnes, F. R.; Jack, M. A. (1997). Weighted matching algorithms and reliability in noise cancelling by spectral subtraction. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP97)*, vol. 2, pp. 1171–1174, München; Germany.