ADVANCED SPEECH ENHANCEMENT WITH PARTIAL SPEECH RECONSTRUCTION

Patrick Hannon, Mohamed Krini, Ingo Schalk-Schupp

Nuance Communications Deutschland GmbH
Acoustic Speech Enhancement Research
Soeflinger Strasse 100, 89077 Ulm, Germany

ABSTRACT

An advanced speech enhancement algorithm is proposed, which employs partial speech reconstruction of highly disturbed speech. The speech reconstruction algorithms assume the source-filter model of speech production and construct estimates of clean speech source and filter signals using features extracted from noisy input. A nonlinear harmonic regeneration scheme for source signals is presented followed by two methods for the estimation of the vocal tract filter characteristics. The quantization method applies a priori trained codebooks using clean speech training data and the parametric estimation method assumes a parabolic continuation of low frequency envelope values. The predicted speech quality of the enhanced speech output is assessed with composite objective measures, while the accuracy of the spectral envelope estimations is analyzed with the log-spectral distance over four manually generated signal-to-noise ratio scenarios.

***Index Terms*—** speech reconstruction, model based estimation, speech enhancement, noise reduction.

1. INTRODUCTION

The limitations of modern speech enhancement algorithms become apparent when basic a priori assumptions of the methods are not fulfilled at runtime. It has been shown that conventional speech enhancement methods perform adequately in environments with medium to high signal-to-noise ratios (SNRs) and relatively stationary background noise [1, 2]. The fact that the performance of conventional speech enhancement is often unsatisfactory at lower SNRs is the major motivation for exploring speech reconstruction systems.

This is where model based approaches to speech enhancement find a high level of applicability. However, there are two main challenges: i) finding a sufficiently good speech model from which to reconstruct speech and ii) establishing a set of features that can be estimated robustly from noisy speech [3].

The work presented here assumes the source-filter model of speech production as the basis model. Thus, it is necessary to develop methods for estimating clean versions of the source and filter representations from the noisy input speech signal. A method of harmonic regeneration in the source signal is presented in section 3.1. For the filter—or spectral envelope—estimation, two methods are described in sections 3.2.1 and 3.2.2. Finally, the construction of a synthetic speech spectrum and the adaptive combination with the noise reduced input spectrum are described in section 3.3.

2. CONVENTIONAL SPEECH ENHANCEMENT

Conventional speech enhancement algorithms aim to reduce background noise while preserving the undisturbed speech signal. Assuming additive noise, the time domain microphone signal, $y(t)$, is represented by

$$y(t) = s(t) + b(t), \tag{1}$$

where $s(t)$ and $b(t)$ are speech and noise respectively.

The Discrete Fourier Transform (DFT) of the n^{th} block of the 16 kHz microphone signal using a Hann window, $\boldsymbol{h}$, of length $M = 512$ is defined as

$$Y_\mu(n) = \sum_{m=0}^{M-1} y(nr - m)\, h_m\, e^{-j2\pi\mu m/M}, \tag{2}$$

with frameshift $r = 128$ and subband index $\mu = 0, \ldots, M$–1.

The Wiener filter algorithm attempts to suppress noise with an adaptive filter, $G_\mu(n)$, applied in the frequency domain as

$$\widehat{S}_{\text{NR},\mu}(n) = Y_\mu(n)\, G_\mu(n)\,, \tag{3}$$

resulting in the frequency domain representation of the clean speech estimate, $\widehat{S}_{\text{NR},\mu}(n)$. Improvements to the Wiener filter algorithm incorporate information from previous frames into the speech enhancement process. In one such method known as recursive Wiener filtering [4], $G_\mu(n)$ is computed as

$$G_\mu(n) = \max\left\{ G_{\min}, 1 - \frac{\hat{\Phi}_{bb,\mu}(n)}{G_\mu(n-1)\cdot|Y_\mu(n)|^2} \right\}, \tag{4}$$

where $\hat{\Phi}_{bb,\mu}(n)$ represents the estimated noise power spectral density [5] and $G_{\min}$ is the maximum attenuation [6].

Although conventional speech enhancement methods perform adequately in environments with SNR $> 10\,$dB, problems arise when the noise masks the desired speech signal. As a consequence, the Wiener filter inadvertently attenuates masked speech simultaneously with the undesired noise.

3. PARTIAL SPEECH RECONSTRUCTION

For enhancing such highly distorted speech signals, a computationally efficient partial speech reconstruction algorithm

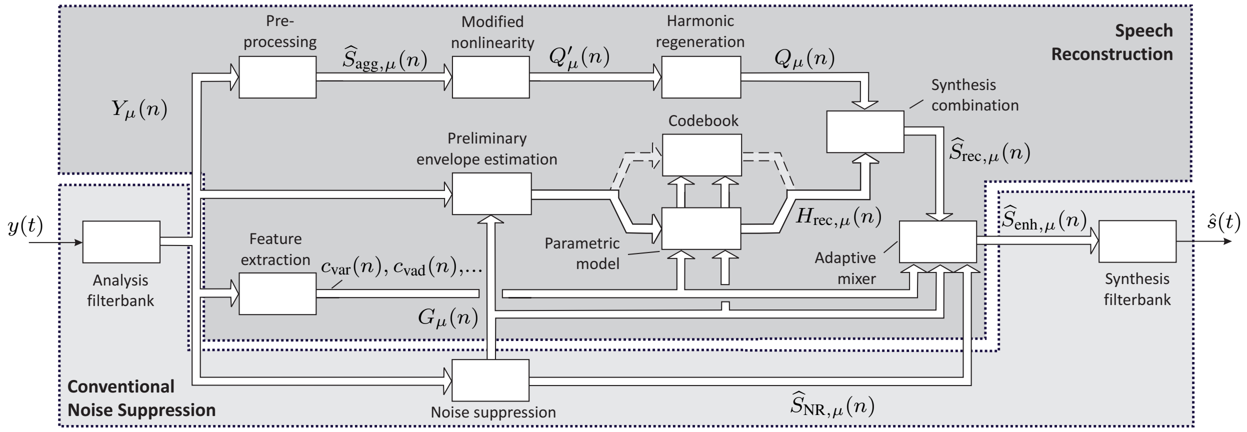

Fig. 1: Proposed speech enhancement algorithm consisting of speech reconstruction, noise suppression, and adaptive mixing.

can be used. The signal flow diagram of the proposed method resulting in an estimate of the undisturbed speech signal $\hat{s}(t)$ is shown in figure 1. The speech signal enhancement consists mainly of three algorithmic parts: conventional noise suppression, partial signal reconstruction, and a mixing unit that adaptively combines both of these signals.

The synthetic speech signal is generated by exploiting correlations with time-frequency regions of speech that exhibit sufficient SNR, using a priori trained models as well as relevant features extracted from the noisy speech signal. In order to achieve a satisfactory speech synthesis, a minimum amount of speech signal portions should still have high SNR levels. This condition is satisfied for automotive applications in most cases where the additive noise components (engine, wind noise, etc.) are concentrated in lower frequencies.

3.1. Harmonic Reconstruction

The first step of the speech reconstruction algorithm is constructing an estimate of the source – or excitation – signal based on an aggressive version of the noise suppression in (3). A high maximum attenuation (e.g. $G_{\text{min}} = -30\,\text{dB}$), resulting in coefficients $G_{\text{agg},\mu}(n)$, is suggested to generate a noise-free signal:

$$\widehat{S}_{\text{agg},\mu}(n) = Y_{\mu}(n)\, G_{\text{agg},\mu}(n)\,. \tag{5}$$

The outcome is used as a reference signal for reconstructing the corrupted speech excitation signal.

The application of a computationally efficient nonlinear operator to $\widehat{S}_{\text{agg},\mu}(n)$ is proposed to reconstruct distorted harmonics. It is well known that by applying a nonlinear characteristic to a harmonic signal, sub- and super-harmonics are produced. However, a full nonlinear characteristic can only be efficiently applied in the time-domain while speech enhancement algorithms are often performed in the frequency or subband domain.

A quadratic characteristic is chosen as the nonlinear operator, which corresponds to the convolution of the signal with itself in the frequency domain. By applying the autoconvolution to a subset of subband signals at lower frequencies only, harmonics in the desired frequency range are regenerated efficiently:

$$Q'_{\mu}(n) = \sum_{m=0}^{M'-1} \widehat{S}_{\text{agg},m}(n)\, \widehat{S}_{\text{agg},\mu+M-1-m}(n). \tag{6}$$

M' denotes the upper frequency bin of the convolution, chosen out of the interval corresponding to frequencies of $1000\,\text{Hz}$–$2000\,\text{Hz}$. Due to the modified nonlinear operator, the harmonics are regenerated at desired frequencies but with biased amplitudes (including components around 0 Hz that must be removed). To reduce this effect, the produced signal is normalized by a smoothed estimate of its magnitude spectral envelope:

$$Q_{\mu}(n) = \frac{Q'_{\mu}(n)}{H_{Q',\mu}(n) + \epsilon} \tag{7}$$

where ϵ is a constant to avoid division by 0. The frequency smoothing of the magnitude spectrum is performed utilizing a first-order IIR filter, first in the positive frequency direction:

$$H'_{Q',\mu}(n) \;=\; \begin{cases} \big|Q'_{\mu}(n)\big|, & \text{if } \mu = 0, \\ \lambda_{\text{f}}\, H'_{Q',\mu-1}(n) + \big(1-\lambda_{\text{f}}\big)\, \big|Q'_{\mu}(n)\big|, & \text{else}, \end{cases} \tag{8}$$

and then in the negative frequency direction in an analogous way leading to $H_{Q',\mu}(n)$. To guarantee stable operation, the constant λ_{f} should be chosen out of the interval $0 \leq \lambda_{\text{f}} < 1$ and $\lambda_{\text{f}} = 0.7$ was utilized for (7).

3.2. Magnitude Spectral Envelope

Having assumed the source-filter model of speech production and given the excitation signal estimation above, it remains to find an appropriate estimate of the ideal vocal tract filter, $H_{S,\mu}(n)$. The observable vocal tract filter for block n of the input signal is represented here by an approximation of the magnitude spectral envelope, $H_{Y,\mu}(n)$, which is extracted from the magnitude of the input spectrum, $|Y_{\mu}(n)|$. For this work, the cost efficient method of spectral envelope extraction presented in (8) is applied with $\lambda_\text{f} = 0.8$. The next sections describe methods for finding estimates of $H_{S,\mu}(n)$ based on features extracted from $H_{Y,\mu}(n)$.

3.2.1. Envelope Estimation Based on Codebooks

The codebook—or quantized—estimate of the spectral envelope, $H_{\text{cb},\mu}(n)$, is derived from an a priori trained codebook, $\boldsymbol{C}^{(K)}_{P+M}$, with $K = 512$ codebook entries. The k^th entry consists of a supervector comprising a log Mel spectrum feature vector of length $P = 24$ followed by the corresponding spectral envelope of length M.

This work includes the log Mel feature vectors to reduce the computational effort of codebook entry selection, which was previously performed using the M-length spectral envelope directly. For the construction of feature vectors, P normalized, triangular Mel filters are applied to the appropriate normalized spectral envelope, noisy or noise reduced. This is performed up to the frequency bin of interest during voiced speech, such as in (6), corresponding here to a frequency of 2000 Hz, and followed by the logarithm operator.

As in previous work [7], normalized codebook entries are extracted from the clean training data set and assigned using the Linde-Buzo-Gray clustering algorithm. The log Mel distance is used during codebook training and at runtime to find the codebook entry most similar to the input feature vector, $\boldsymbol{c}_\text{lms}(n)$. This distance measure is defined as

$$D^{(k)}_\text{lm}(n) = \sum_{p=0}^{P-1} G_{\text{mel},p}(n) \left(C^{(k)}_p - c_{\text{lms},p}(n) \right)^2 , \qquad (9)$$

for $k = 0, \ldots, K-1$, and $G_{\text{mel},p}$ are the Mel filtered Wiener filter coefficients that serve as SNR dependent weights. During training this SNR dependent weighting factor is not applicable. The index of the entry with the smallest distance is calculated according to

$$\kappa(n) = \underset{0 \le k < K}{\arg\min}\, D^{(k)}_\text{lm}(n). \qquad (10)$$

Finally, the codebook estimate of the magnitude spectral envelope is assigned to the entry with the smallest distance

$$H_{\text{cb},\mu}(n) = C^{(\kappa(n))}_{P+\mu}. \qquad (11)$$

3.2.2. Envelope Estimation Based on a Parametric Model

As an alternative to the codebook method, a novel parametric model for low frequency envelope estimation, represented by $H_{\text{par},\mu}(n)$, is proposed [8]. The model is based on a logarithmic parabola shape that extends towards low frequencies from the lowest frequency bin, $\mu_\text{fix}(n) = \mu$, where $G_\mu(n) > \gamma_\text{par}$.

As seen in figure 2, the logarithmic parabola shape is anchored to the reference envelope $H_{\text{avg},\mu}(n)$ defined by

$$H_{\text{avg},\mu}(n) = \frac{H_{Y,\mu}(n) + H_{\widehat{S}_\text{NR},\mu}(n)}{2}, \qquad (12)$$

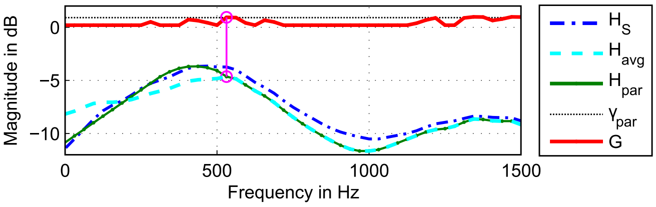

Fig. 2: Example illustration of parabolic envelope reconstruction

the average of the noisy and noise reduced input envelopes, so that

$$H_{\text{par},\mu}(n) = H_{\text{avg},\mu}(n) \quad \forall \mu \ge \mu_\text{fix}(n). \qquad (13)$$

At that point the curvature $J(n)$, the initial exponential slope $m_\text{init}(n)$, the minimum exponential slope $m_{\text{min},\mu}(n)$, and the maximum exponential slope $m_\text{max}(n)$, of the desired parabola are computed from signal features and the logarithmic parabola is recursively extrapolated towards lower frequencies using:

$$H_{\text{par},\mu}(n) = \frac{J(n) \cdot H_{\text{par},\mu+1}(n)}{\min\left\{m_\text{max}(n), \max\left\{m_{\text{min},\mu}(n), m_\mu(n)\right\}\right\}} \qquad \forall\, \mu < \mu_\text{fix}(n), \qquad (14)$$

where

$$m_\mu(n) = \begin{cases} m_\text{init}(n), & \text{if } \mu = \mu_\text{fix}(n) - 1, \\ \dfrac{H_{\text{par},\mu+2}(n)}{H_{\text{par},\mu+1}(n)}, & \text{else.} \end{cases} \qquad (15)$$

The parabola's minimum exponential slope is set to the local exponential slope of the noisy signal's envelope:

$$m_{\text{min},\mu}(n) = \frac{H_{Y,\mu+1}(n)}{H_{Y,\mu}(n)}. \qquad (16)$$

To find good parameter estimators, the training data set is used to find the parameters of the best-fitting parabola in each frame with reliably detectable voice activity. This is done using a simple minimum search on the log-spectral distance (LSD) measure (26) between estimator and the known clean speech signal's envelope $H_{S,\mu}$. Also, each frame's signal features are calculated. The resulting optimal parameters are then stored together with the corresponding features in a supervector. Approximate functional dependencies between features and parameters are heuristically fit to these data separately for each noise condition, introducing coefficients a_i and b_i for $i \in \{m_{\mu_\text{fix}}, J, m_\text{max}\}$.

In previous experiments [8], the most useful feature proved to be the reference envelope's exponential slope around $\mu_\text{fix}(n)$:

$$m_{\text{avg},\mu_\text{fix}(n)}(n) = \sqrt{\frac{H_{\text{avg},\mu_\text{fix}(n)+1}(n)}{H_{\text{avg},\mu_\text{fix}(n)-1}(n)}}. \qquad (17)$$

This feature is used to estimate the parabola parameters via the fit coefficients determined in the training:

$$m_{\mu_{\text{fix}}(n)}(n) = a_{m_{\mu_{\text{fix}}}} \cdot m_{\text{avg},\mu_{\text{fix}}(n)}(n) + b_{m_{\mu_{\text{fix}}}} \tag{18}$$

$$J(n) = a_J \cdot m_{\text{avg},\mu_{\text{fix}}(n)}(n) + b_J \tag{19}$$

$$m_{\max}(n) = a_{m_{\max}} \cdot m_{\text{avg},\mu_{\text{fix}}(n)}(n) + b_{m_{\max}}. \tag{20}$$

3.3. Synthesis Combination and Adaptive Mixer

The excitation and envelope estimates are combined:

$$\widehat{S}'_{\text{rec},\mu}(n) = Q_\mu(n) \cdot H_{\text{rec},\mu}(n), \tag{21}$$

where
$$H_{\text{rec},\mu}(n) \in \{H_{\text{par},\mu}(n), H_{\text{cb},\mu}(n)\}. \tag{22}$$

Either an SNR-dependent mixture of the reconstructed and the noise reduced signals is performed or an additional attenuation, G_{att}, of the noise reduced spectrum is applied:

$$\widehat{S}_{\text{rec},\mu}(n) \tag{23}$$
$$= \begin{cases} \big(1-\alpha_\mu(n)\big)\,\widehat{S}'_{\text{rec},\mu}(n) + \alpha_\mu(n)\,\widehat{S}_{\text{NR},\mu}(n), \\ \quad \text{if } \left(\left|\widehat{S}'_{\text{rec},\mu}(n)\right| > K_1 \left|\widehat{S}_{\text{NR},\mu}(n)\right|\right) \wedge \left(c_{\text{var}}(n) > K_2\right), \\ G_{\text{att}}\,\widehat{S}_{\text{NR},\mu}(n), \quad \text{else.} \end{cases}$$

Heavily distorted harmonics are detected if the synthesized magnitude spectrum exceeds the noise reduced magnitude spectrum by a specific margin. For enhanced separability between voiced speech and noise or noise-like periods the variance measure, $c_{\text{var}}(n)$, of the subband SNR in dB over the frequency within a certain range of values (e.g. 200 Hz–1500 Hz) are taken into account. Further measures like the statistical voice activity detection, $c_{\text{vad}}(n)$, from [9] can be employed additionally. The value of K_1 corresponds to 6 dB and $K_2 = 25$ dB. For noise-only subband signals, an additional attenuation G_{att} is applied. The mixing weights can be normalized within the range of $[0, 1]$:

$$\alpha_\mu(n) = \frac{\min\left\{\gamma_{\text{rec}},\, G_\mu(n)\right\} - G_{\min}}{\gamma_{\text{rec}} - G_{\min}}. \tag{24}$$

Once the synthetic speech signal is generated, it is adaptively combined with the conventionally noise reduced signal from (3) in the subband domain according to:

$$\widehat{S}_{\text{enh},\mu}(n) \tag{25}$$
$$= \begin{cases} \widehat{S}_{\text{NR},\mu}(n), & \text{if } \left(G_\mu(n) > \gamma_{\text{rec}}\right) \vee \left(\mu > \mu_{\text{r, max}}\right), \\ \widehat{S}_{\text{rec},\mu}(n), & \text{else.} \end{cases}$$

The adaptive mixing is only utilized up to a predefined frequency bin, $\mu_{\text{r, max}}$, chosen from the same interval as M' in (6). In good SNR conditions, i.e. if the SNR-dependent estimate $G_\mu(n)$ exceeds a predefined threshold $\gamma_{\text{rec}} > G_{\min}$, the conventionally noise suppressed output is used.

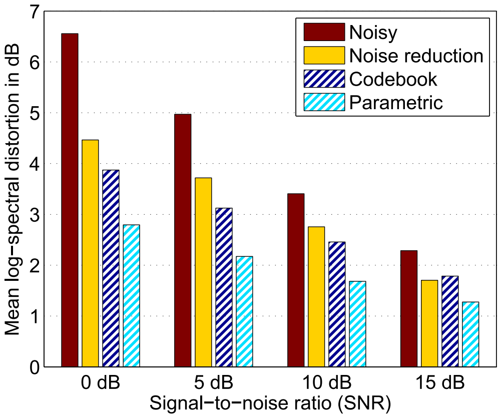

Fig. 3: Log-spectral distortion of envelope estimates

4. EVALUATIONS

A test set of 22 speech files per each SNR $\in \{0, 5, 10, 15\}$ dB is generated for the evaluation using different car impulse responses and stationary background noises. Both the noise and the impulse responses were measured in an automotive acoustic environment. The maximum noise attenuation was set to $G_{\min} = -14$ dB.

4.1. Objective Envelope Estimation Performance

The LSD is used for envelope estimation evaluation:

$$D_{\text{ls}} = \frac{10}{L}\sum_{n=0}^{N}\sqrt{\sum_{\mu=0}^{\mu_{\text{r,max}}} \frac{K_{\mu,n}}{\overline{K}_n}\,\lg^2\left\{\frac{\max\{H_{Y,\mu}(n), \delta_Y\}}{\max\{H_{\text{rec},\mu}(n), \delta_{\text{rec}}\}}\right\}}. \tag{26}$$

The lower bound is defined as

$$\delta_Y = 10^{-5}\max_{\mu,n}\{H_{Y,\mu}(n)\}, \tag{27}$$

and similarly for δ_{rec}. For evaluating the envelope distortion, the binary mask $K_{\mu,n} \in \{0, 1\}$ selects only those components that satisfy the condition: $H_{Y,\mu}(n) \geq \delta_Y$. The corresponding normalization is given by $\overline{K}_n = \max\{\sum_\mu K_{\mu,n}, 0.1\}$. N represents the number of potential frames and L corresponds to the number of frames for which $\overline{K}_n \geq 1$. The analysis of the distortion is only performed at lower frequencies up to 1200 Hz.

The evaluation results as seen in figure 3 show that both the codebook, and the parametric estimator provide better results than the noisy or noise reduced envelopes, except for a very good SNR of 15 dB. The parametric approach outperforms the other methods in all of the noise conditions.

4.2. Pseudo-Subjective Composite Results

In [10], several composite objective measures are proposed that combine individual objective measures to predict the subjective quality of noise reduced signals after the application of speech enhancement algorithms. The subjective quality ratings were obtained using the ITU-T P.835 methodology designed to evaluate the quality of enhanced speech along three

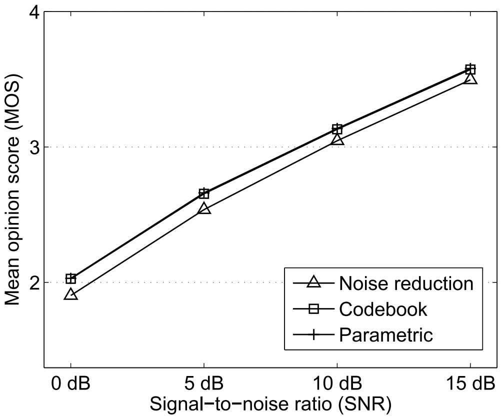

Fig. 4: Signal quality composite score

dimensions: signal distortion, noise distortion, and overall quality.

As seen in figure 4, the proposed speech reconstruction method improves the composite mean opinion score (MOS) for signal distortion compared to that of conventional noise reduction. Both methods for spectral envelope estimation perform equally well and this suggests a heavy dependence on the excitation signal generation algorithm. The composite measure for noise distortion (not pictured) suggests improvements only in the 0 dB SNR scenario, and thus the improvement in overall quality is slightly less than that of signal distortion.

5. CONCLUSIONS

Complementing conventional speech enhancement with the proposed partial speech reconstruction algorithms achieves improvements in both signal and noise distortions. Based on the analyses presented above, it can be concluded that corrections of the distorted spectral envelopes are possible to a high degree at all SNR levels, especially in the case of the parametric parabola estimation. This envelope correction has the ability to improve the quality of distorted speech given an excitation signal of adequate quality.

The predicted speech distortion improvement from the composite objective measures shows the same improvement for both the parametric and the codebook based envelope estimation methods when combined with the harmonic source regeneration method. This result suggests that the excitation signal has a large influence on the perceived quality of the enhanced speech output and that there may exist a subjective tolerance for errors in spectral envelope estimation.

An example of the enhanced speech output can be seen in figure 5, where spectra of the noisy microphone signal, conventional noise reduction signal, and the partially reconstructed signal are depicted side by side. The presence of harmonic spectral lines at low frequencies is plainly visible here and this effect contributes to a more natural sounding speech output after the application of speech reconstruction.

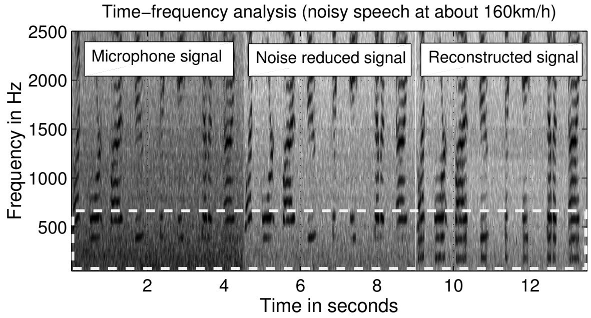

Fig. 5: Analyses of speech originating in an automotive environment

6. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on ASSP*, vol. 32, no. 6, pp. 1109–1121, 1984.

[2] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, The MIT Press, 1964.

[3] P. Harding and B. Milner, "Speech enhancement by reconstruction from cleaned acoustic features," in *Interspeech Proceedings*, 2011, pp. 1189–1192.

[4] K. Linhard and T. Haulick, "Spectral noise subtraction with recursive gain curves," in *ICSLP Proceedings*, 1998, pp. 1479–1482.

[5] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. on SAP*, vol. 11, no. 5, pp. 466–475, 2003.

[6] M. Krini and G. Schmidt, "Model-based speech enhancement for automotive applications," in *ISPA Proceedings*, 2009, pp. 632–637.

[7] P. Hannon and M. Krini, "Spectro-temporal features for excitation signal quantization in a speech reconstruction system," in *DSP Workshop for In-Vehicle Systems*, 2011.

[8] I. Schalk-Schupp, "Improved noise reduction for handsfree communication in automobile environments," diploma thesis, O.-v.-Guericke-Univ. Magdeburg, 2012.

[9] P. Loizou, *Speech enhancement: Theory and Practice*, CRC, Boca Raton, 2007.

[10] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. on SAP*, vol. 16, no. 1, pp. 229–238, 2008.