Improved CEM for speech harmonic enhancement in single channel noise suppression

Yanjue Song and Nilesh Madhu
IDLab, Ghent University - imec, Ghent, Belgium

***Abstract*—The periodic nature of voiced speech is often exploited to restore speech harmonics and to increase inter-harmonic noise suppression. In particular, a recent paper proposed to do this by manipulating the speech harmonic frequencies in the cepstral domain. The manipulations were carried out on the cepstrum of the excitation signal, obtained by the source-filter decomposition of speech. This method was termed Cepstral Excitation Manipulation (CEM). In this contribution we further analyse this method, point out its inherent weakness and propose means to overcome it. First of all, it will be shown by both illustrative examples and theoretical analysis that the existing method underestimates the excitation, especially at low signal to noise ratio (SNR) conditions. This inherent weakness leads to speech harmonic weakening and vocoding due to the insufficient noise suppression in the inter-harmonic regions. Then, we propose two modifications to improve the robustness and performance of CEM in low SNR cases. The first modification is to use an instantaneous amplifying factor adapted to the signal, instead of a pre-defined constant, for the excitation cepstrum. The second modification is to smooth the excitation cepstrum to preserve additional fine structure, instead of discarding it. These modifications result in better preservation of speech harmonics, more refined fine structure and higher inter-harmonic noise suppression. Experimental evaluations using a range of standard instrumental metrics conclusively demonstrate that our proposed modifications clearly outperform the existing method, especially in extremely noisy conditions.**

***Index Terms*—speech enhancement, harmonic synthesis, cepstral smoothing, cepstral excitation manipulation.**

I. Introduction

Single channel speech enhancement is widely used in several communications applications to suppress background noise and improve speech quality. Many efforts have been made in the topic and this is still an active field of research. The majority of statistical speech enhancement algorithms apply a gain function to the short time-frequency representation of speech. Based on the estimation of the power spectral density of the noise and the underlying speech, the so-called *a priori* SNR and *a posteriori* SNR are, typically, the two crucial parameters required for the gain function calculation.

Under individual independent Gaussian assumptions of noise and speech spectral coefficients, Ephraim and Malah derived the classical Minimum-Mean Square Error Short-Time Spectral Amplitude (MMSE-STSA) estimator [1] and Minimum-Mean Square Error Log-Spectral Amplitude (MMSE-LSA) estimator [2], where the *a priori* SNR is estimated by the *Decision-Directed* (DD) approach. One well-known drawback of MMSE estimator is musical noise in the enhanced audio. The recursive smoothing of the DD approach tackles it to some degree, but also leads to speech distortions. This drawback has been handled in various ways in the literature. For example, [3] proposed the Two-Step Noise Reduction (TSNR) technique to not only solve the one-frame delay introduced by the bias of *a priori* SNR estimation but also combined it with *harmonic regeneration noise reduction* (HRNR) to restore the distorted harmonics by applying a non-linearity in the time domain. Cepstro-temporal smoothing has been proposed in [4] to reduce the musical tones. It essentially applies a post-filter to the MMSE gain estimate in the cepstrum domain. The recursive smoothing of high-order cepstral coefficients of the gain function reduces abrupt changes in its fine structure, thereby lowering musical noise. This technique can also be applied to the cepstral coefficients of the enhanced spectrum directly. For instance, a selective cepstro-temporal smoothing scheme was proposed in [5]. Typically, these approaches target a better enhancement of the speech harmonics in voiced speech regions.

On the other hand, the speech production process can be abstracted as a source-filter model that captures the harmonic structure in speech with low mathematical complexity [6]. Thereby, the speech signal is decomposed into an envelope and an excitation signal. There have been analysis-synthesis speech enhancement frameworks based on this decomposition, leading to learning-based envelope enhancement solutions, e.g., [7]–[9]. In addition to the envelope enhancement, a synthetic excitation signal is also introduced in [8], [9], to address the musical noise problem. But it has also been reported that synthetic speech lacks naturalness. For this reason, and since the source-filter decomposition is sensitive to noise, the approaches are usually combined with the aforementioned spectral amplitude estimators, resulting in a TSNR framework.

More recently, there have been attempts to improve speech quality by directly manipulating the excitation signal. In [10], [11], the cepstral representation of the excitation signal is adopted to highlight the periodic structure of speech. Using this representation, it is possible to get a clearer speech estimation and even to restore low-amplitude or lost harmonics by amplifying the periodic structure of voiced speech in the cepstral domain. Consequently, this approach is termed Cepstral Excitation Manipulation (CEM). A benchmark of CEM against the relevant state-of-the-art in [10] indicated the potential of this approach to further improve the enhanced speech quality. This method was, further, combined with envelope estimation methods in [11], thereby integrating the benefits of envelope and excitation improvement for speech enhancement. This demonstrated that in addition to being used

as a stand-alone component, CEM could also be piggy-backed onto other speech enhancement frameworks – which makes this idea versatile.

In [10], the excitation of speech is replaced by either an idealised, synthetic excitation or by selecting one from pre-trained cepstral excitation templates. The non-template method has the advantage that it is capable of recovering lost harmonics in a relatively simple manner: all but two of the cepstral coefficients of the excitation signal are set to zero (the process termed 'cepstral nulling'). The zeroth coefficient, corresponding to the energy term, is preserved, along with the cepstral coefficient that has maximum amplitude within the range of allowed speech fundamental frequencies. This latter coefficient is further amplified by a pre-determined constant, thereby implicitly emphasising the spectral peaks at the fundamental frequency and its harmonics. However, as discussed in [10], this method shares the disadvantage of the lack of naturalness in the enhanced audio with the analysis-synthesis framework. This is due to the fact that most of the cepstrum has been discarded, which means the loss of speech fine structure. The amendment provided by [10] is to introduce cepstral excitation templates in either a speaker-dependent or a speaker-independent manner. However, this solution adds to the complexity of the system because of the template training and extra models required. Another problem of this method comes from the constant pitch amplifying factor. When the harmonics are corrupted by noise, the dynamic range of the boosted speech excitation signal is insufficient with a pre-determined amplifying factor. In such cases, the noise at the inter-harmonic frequencies is poorly suppressed, which results in a vocoder-like effect. Thus, classical CEM fails to maintain the sharpness of harmonics in the enhanced output signals, especially in low SNR conditions.

In this contribution, we focus on the above shortcomings of CEM and propose means to address them. First, instead of using a fixed pitch amplifying factor, the dynamic range of the synthetic excitation is instantaneously and adaptively estimated based on the input signal spectrum. Second, the excitation signal cepstrum is selectively averaged along the quefrency, which preserves the spectral fine structure better and results in a more natural estimate of the underlying speech spectrum. This can serve as an alternative to speaker excitation templates trained on large corpora as in [10]. Lastly, these two improvements can be combined to yield a robust, improved CEM approach with better performance in all SNR conditions.

The paper is structured as follows: the baseline CEM method, together with the speech enhancement framework, is introduced in Section II. Representative examples are first shown in Section III to demonstrate the weakness of classical CEM in excitation synthesis and its consequence. Our modifications aimed at addressing these shortcomings are next presented. The proposed methods are thoroughly evaluated in Section IV and the conclusions are presented in Section V.

II. Cepstral excitation manipulation (CEM) baseline

We consider an additive mixture of an underlying speech signal $s(n)$ mixed with the background noise signal $v(n)$. The goal of speech enhancement is to get a clean speech estimate $\widetilde{s}(n)$ of better quality and/or intelligibility given the noisy observation $y(n) = s(n) + v(n)$. With an M-point windowed Fourier Transform, the mixture can be represented in the short-time Fourier Transform (STFT) domain as the summation of the short term spectra of speech S_l and of noise V_l[1]:

$$Y_l(m) = S_l(m) + V_l(m)\,, \tag{1}$$

where l is the frame index and m is the frequency bin index.

The gain function method is adopted in this framework where, for each frame l, the estimate of the underlying clean speech spectrum is obtained by multiplying the noisy input spectrum with a gain function $G_l(m)$:

$$\widetilde{S}_l(m) = G_l(m) Y_l(m)\,. \tag{2}$$

The $G_l(m)$ are typically (see, e.g., [12]) obtained as functions of the following two parameters: the *a priori* SNR $\xi_l(m)$ and the *a posteriori* SNR $\gamma_l(m)$, respectively defined as:

$$\xi_l(m) = \frac{\lambda_{s,l}(m)}{\lambda_{v,l}(m)}\,, \tag{3}$$

and

$$\gamma_l(m) = \frac{|Y_l(m)|^2}{\lambda_{v,l}(m)}\,, \tag{4}$$

where $\lambda_{s,l}(m)$ and $\lambda_{v,l}(m)$ represent the power spectral densities (PSDs) of the speech and noise signals, respectively. However, since the true values of $\lambda_{s,l}(m)$ and $\lambda_{v,l}(m)$ are not available, their estimates, $\widehat{\lambda}_{s,l}(m)$ and $\widehat{\lambda}_{v,l}(m)$, are substituted into the above formulae to compute the *a priori* and *a posteriori* SNRs for the calculation of the gain function.

A. Overview of CEM-based Speech Enhancement Framework

To improve the speech estimate $\widehat{\lambda}_{s,l}(m)$, especially for the voiced speech segments, the two-stage speech enhancement framework is proposed in [10], and is summarised below.

In the first stage, a preliminary noise reduction is applied, resulting in an initial speech estimate $\widehat{S}_l(m)$. This is obtained by applying the MMSE-LSA gain function together with the decision-directed (DD) approach [2]. The noise PSD $\widehat{\lambda}_{v,l}(m)$ is estimated by the Minimum Statistics (MS) approach [13].

Using LPC analysis [6], $\widehat{S}_l(m)$ is decomposed into the *envelope* $\widehat{H}_l(m)$ and the *residual* $\widehat{R}_l(m)$, which is also termed the speech excitation signal.

The key idea of [10] lies in the manipulation of this excitation signal in the *cepstral* domain. First, the excitation signal $\widehat{R}_l(m)$ is converted to cepstrum where fundamental frequency can be easily detected. The speech harmonics are selectively boosted as detailed below. Applying the original speech envelope $|\widehat{H}_l(m)|$ to this enhanced excitation signal $|\overline{R}_l(m)|$, an idealised speech estimate $|\overline{S_l}(m)|$ can be obtained. A new *a priori* SNR $\widetilde{\xi}_l$ is then calculated from this harmonic-enhanced estimate $|\overline{S_l}(m)|$ to obtain the final gain function

[1]We follow the standard convention: uppercase letters indicate quantities in the spectral domain; lowercase variables are time-domain signals. Since the cepstrum, defined as the inverse transform of the logarithmic spectrum, is also a quasi-temporal representation and lowercase letters are adopted for cepstral variables as well.

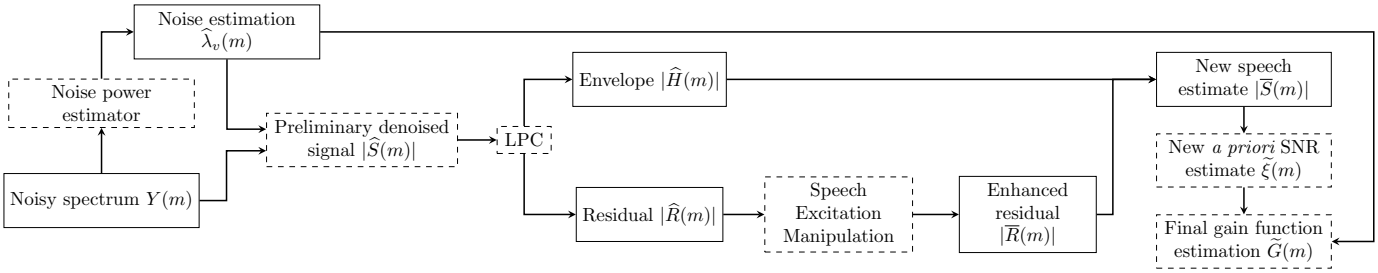


Fig. 1: Block diagram of the gain function calculation in two-stage noise reduction. Dashed boxes represent manipulation blocks whereas solid rectangular boxes indicate data contained. Please note that all terms are in the STFT domain, where the frame index l has been dropped for conciseness.

for that frame. This method is named as CEM_{ID} because it replaces the original excitation with an idealised one. Along with the traditional MMSE-LSA approach, it forms a baseline for our work. The signal flow graph of CEM_{ID} is graphically illustrated in Fig. 1. The details of the cepstral manipulation of CEM are now presented.

B. CEM_{ID} in detail: F_0 Detection

To analyse the periodicity of voiced frames, the excitation signal amplitude $|\widehat{R}_l(m)|$ is transformed into the cepstral domain by a Q-point Discrete Cosine Transformation of type II (DCT-II):

$$\begin{aligned} c_l(q) &= \text{DCT}\{\ln(|\widehat{R}_l(m)|)\} \\ &= \sum_{m=0}^{M/2} \ln(|\widehat{R}_l(m)|) \cdot \cos\left[\frac{\pi q}{Q}(m+0.5)\right], \end{aligned} \tag{5}$$

where $q = \{0, 1, \ldots, Q-1\}$ denotes the quefrency bin index. Since the amplitude spectrum is symmetric, only half of the residual spectrum $|\widehat{R}_l(m)|$ is required for cepstrum calculation and Q is set to $M/2+1$. The fundamental frequency and its harmonics correspond to a peak in the cepstrum. For the signal sampled at f_s, the relationship between frequency f and its corresponding quefrency bin is $f = f_s/q$. Therefore, the fundamental frequency F_0 at frame l can be obtained by finding its corresponding quefrency bin q_{F_0} where the excitation cepstrum $c_l(q)$ achieves its maximum in the allowed quefrency range. The estimated fundamental frequency is then given by

$$F_0(l) = \frac{f_s}{q_{F_0}(l)}, \tag{6}$$

where

$$q_{F_0}(l) = \underset{q \in \mathcal{Q}}{\text{argmax}}\{c_l(q)\}. \tag{7}$$

Given that the fundamental frequency of human speech usually falls in the range from 50 to 500 Hz [6], the search boundary in the quefrency domain is constrained correspondingly as $\mathcal{Q} = [q_{f=500}, \ldots, q_{f=50}]$.

C. CEM_{ID} in detail: Excitation Manipulation

Following the identification of $q_{F_0}(l)$, the original excitation is completely replaced by a synthesised excitation $\overline{c}_l(q)$: $c_l(0)$, as an indication of energy level, is preserved in the idealised excitation $\overline{c}_l$. The cepstral peak is *amplified* by a pre-determined constant $\alpha_c (> 1)$. The rest of the cepstrum is discarded, i.e.,

$$\overline{c}_l(q) = \begin{cases} c_l(0), & q = 0 \\ \alpha_c \cdot c_l(q), & q = q_{F_0},\ \alpha_c > 1 \\ 0, & \text{otherwise} \end{cases}. \tag{8}$$

This excitation manipulation helps with speech enhancement in two ways. On the one hand, the speech harmonics are emphasised by scaling up the harmonic-related cepstral peak. On the other hand, the remaining noise is removed by nulling the excitation cepstrum.

The new speech residual amplitude $|\overline{R}_l(m)|$ spectrum can be acquired by the inverse DCT-II (iDCT) of the idealised excitation cepstrum $\overline{c}_l$:

$$\begin{aligned} |\overline{R}_l(m)| &= \exp(\text{iDCT}\{\overline{c}_l\}) \\ &= \exp\left(\frac{\overline{c}_l(0)}{Q} + \frac{2}{Q}\sum_{q=1}^{Q-1} \overline{c}_l(q) \cdot \cos\left[\frac{\pi q}{Q}(m+0.5)\right]\right). \end{aligned} \tag{9}$$

The synthesis procedure described by Eq. (9) could lead to false peaks or an undesired rising tendency at the edges of the spectrum $|\overline{R}_l(m)|$. This is addressed by a cosine decay in [10] at the spectral edges to avoid these artefacts in the enhanced signal, namely by linearly extending the spectrum from the trough before the first peak and from the trough following the peak of the last harmonic to the respective edges.

D. Speech Estimation

With the idealised excitation $|\overline{R}_l(m)|$ and the speech envelope $|\widehat{H}_l(m)|$, the speech amplitude spectrum is synthesised as $|\overline{S_l}(m)| = |\overline{R}_l(m)| \cdot |\widehat{H}_l(m)|$. Instead of directly using this synthetic spectrum as clean speech estimate, it is proposed to use this idealised spectrum to *re-estimate* the *a priori* SNR $\widetilde{\xi}_l(m)$ as:

$$\widetilde{\xi}_l(m) = \frac{|\overline{S_l}(m)|^2}{\widehat{\lambda}_{v,l}(m)}. \tag{10}$$

Using this $\widetilde{\xi}_l(m)$ and the previously computed *a posteriori* SNR $\widehat{\gamma}_l(m)$, the final *gain* function $\widetilde{G}_l(m)$ is computed in the standard manner (e.g., MMSE-LSA). Applying this gain function yields the final clean speech spectrum estimate: $\widetilde{S}_l(m) = \widetilde{G}_l(m) Y_l(m)$. The clean speech estimate in the time domain is obtained by applying iDFT and overlap-add synthesis.

III. Improved Excitation Manipulation

A. Analysis of the drawbacks of CEM

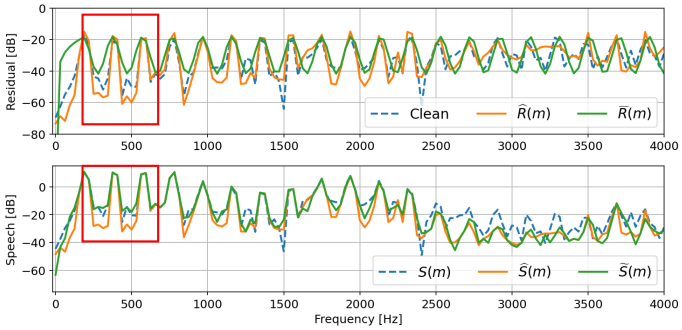
Fig. 2 presents the residual spectrum and speech estimation by CEM_{ID} of a voiced frame corrupted by white noise at SNRs of 15 dB (Fig. 2a) and of -5 dB (Fig. 2b). The speech excitation idealisation is reasonably accurate when SNR $= 15$ dB; however, CEM_{ID} is clearly influenced by residual noise for the same speech segment when SNR $= -5$ dB. Since CEM_{ID} is driven by the excitation manipulation and the envelopes are not modified, the deterioration of CEM_{ID} at -5 dB comes solely from the (insufficient) excitation synthesis. Compared with the clean speech spectrum, the estimated speech spectrum (green curve) at -5 dB (Fig. 2b) is strongly overestimated between the harmonics. From the upper panel, it can be observed that the idealised excitation is even less clear than the excitation spectrum of the preliminary denoised signal (orange curve). As the result, the enhanced spectrum loses the 'sharpness' of its harmonics in voiced frames, especially for the first few harmonics (e.g., the harmonics inside the red square of Fig. 2b). When being used to calculate the new *a priori* SNR, this weakened periodic structure results in poorer inter-harmonic noise suppression, leading to a vocoding effect in the final speech estimate.

To understand this inherent weakness of CEM from an analytic perspective, we take a closer look at the enhanced excitation spectrum, which is written in the log domain as

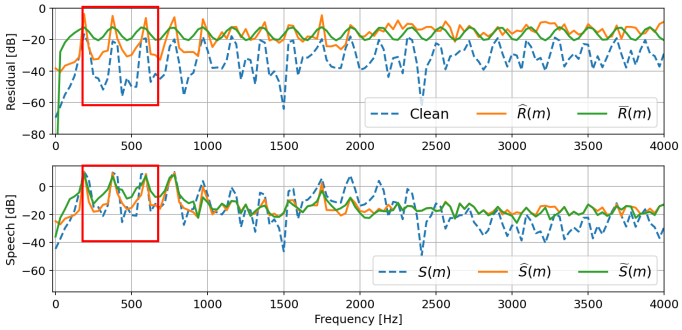
$$\begin{aligned}\ln(|\overline{R}_l(m)|) &= \text{iDCT}\{\overline{c}_l\} \\ &= \frac{\overline{c}_l(0)}{Q} + \frac{2}{Q}\sum_{q=1}^{Q-1}\overline{c}_l(q)\cdot\cos[\frac{\pi q}{Q}(m+0.5)] \\ &= \frac{\overline{c}_l(0)}{Q} + \frac{2}{Q}\overline{c}_l(q_{F_0})\cdot\cos[\frac{\pi q_{F_0}}{Q}(m+0.5)] + \\ &\quad \frac{2}{Q}\sum_{q\in[1,2,\ldots,Q-1]}^{q\neq q_{F_0}}\overline{c}_l(q)\cdot\cos[\frac{\pi q}{Q}(m+0.5)] \\ &= \frac{\overline{c}_l(0)}{Q} + \text{iDCT}\{\mathcal{F}(c_{\text{pitch}})\} + \text{iDCT}\{\mathcal{G}(c_{\text{rest}})\}\end{aligned} \tag{11}$$

where $c_{\text{pitch}} = [0, 0, \ldots, 0, c_l(q_{F_0}), 0, \ldots, 0]$ and $c_{\text{rest}} = [0, c_l(1), c_l(2), \ldots, c_l(q_{F_0}-1), 0, c_l(q_{F_0}+1), \ldots, c_l(q_{F_0})]$. $\mathcal{F}(\cdot)$ and $\mathcal{G}(\cdot)$ are the manipulating functions on the respective cepstrum components. The three terms correspond to the log-spectrum energy, the harmonics, and the fine structure of the excitation. Accordingly, different manipulations are applied to the three terms in CEM_{ID}: the log-spectrum energy term remains untouched, $\mathcal{F}(\cdot)$ amplifies the harmonic term with a constant factor α_c to emphasise the periodic structure, and $\mathcal{G}(\cdot)$ sets the fine structure term to $[0, 0, \ldots, 0]$ for extra noise suppression and to reduce musical noise.

In [10], CEM_{ID} was designed to overestimate harmonic amplitudes by a fixed, pre-determined amplifying factor α_c. However, as demonstrated by the performance degradation of CEM_{ID} from the high-SNR condition to the low-SNR condition in Fig. 2, the α_c suggested by the authors could be insufficient in certain cases. To understand this, recall that cepstral coefficients are obtained by the inner product



(a) With white noise, SNR $= 15$ dB



(b) With white noise, SNR $= -5$ dB

Fig. 2: Speech enhancement by CEM on a voiced frame corrupted by white noise at different SNRs. In each figure, the clean reference, the preliminary denoised result ($\widehat{\cdot}$), the synthesised result ($\overline{\cdot}$), and the final estimate ($\widetilde{\cdot}$) are presented. The *upper panel* of each sub-figure compares the *synthetic excitation* spectrum with that of the preliminary denoised and the clean speech, whereas the lower panel compares the *final clean speech estimate* based on the idealised excitation with that of the preliminary denoised and the clean signal. The method suffers from vocoding, especially in the frequency region delineated by the red square. The artefact results from the amplitude underestimation of the synthetic excitation. This vocoding effect reduces when SNR increases.

of the logarithmic amplitude spectrum and cosine basis functions, computed over the entire spectrum. However, when the harmonics are less pronounced (e.g., due to noise), the excitation signal loses its periodic structure in these regions. In such a case, the cepstral coefficient corresponding to F_0 can still stand out because of the recognisable harmonics (typically in the lower frequency regions); however, its value is attenuated because the harmonic regions are averaged with the unstructured ones. This issue becomes more prominent for wideband speech since harmonic structure is typically more compromised at the higher frequencies regions, leading to more distorted STFT bins that attenuate the F_0-related coefficient. Thus, no matter what *constant* value is chosen as the amplifying factor, there is always the risk of insufficient excitation in CEM_{ID}.

The other problem of CEM_{ID} is that cepstrum nulling leads to artificiality in the enhanced speech, as has been noticed in [10]. Eq. (8) has showed that CEM_{ID} does *not* preserve

any fine spectral structure. It is therefore proposed in [10] to improve the naturalness of the enhanced speech by the template method, where each template is particular to a range of fundamental frequencies. From the perspective of Eq. (11), this modification replaces the nulling function $\mathcal{G}(\cdot)$ of CEM_{ID} by the F_0-related templates which introduces more coefficients into c_{rest}.

By reformulating the speech excitation signal synthesis into three individual, interpretable components in the above analysis, it is easier to understand and appreciate the two limitations of CEM_{ID} when used for speech harmonic enhancement. In the following sections, we will propose our modifications targeting these two shortcomings. We will leave the energy term intact, and enhance the other two terms separately.

B. Residual Amplitude Estimation (RAE)

The first modification is to replace the constant scaling factor α_c of CEM_{ID} with a *data-adaptive* factor. Since the energy of voiced speech is typically concentrated in the low-frequency region, as shown in Fig. 2b, the preliminary noise reduction performs better in the first few low-frequency harmonics, where the SNR is higher. Therefore, we propose to deduce the speech excitation dynamic range from these clearer harmonics. The other consequence of this energy distribution is the decay of the excitation dynamic range as frequency increases. Therefore, *one* amplifying factor that suits low-frequency harmonics will overestimate the high-frequency harmonics and lead to annoying artefacts in the enhanced audio. Motivated by these observations, we propose the residual amplitude estimation (RAE) for $\mathcal{F}(c_{\text{pitch}})$ in Eq. (11). Our RAE consists of three components: an adaptive amplifying factor τ for the excitation dynamic range, an amplitude decay function $\omega(m)$ to avoid overestimation in the high-frequency region, and a cosine function to model the harmonic structures. Based on these three components, the harmonic term of the synthetic excitation can thus be written in the log-domain as:

$$\text{iDCT}\{\mathcal{F}(c_{\text{pitch}})\} = \tau \cdot \omega(m) \cdot \cos[\frac{\pi q_{F_0}}{Q}(m+0.5)]\,. \quad (12)$$

1) Adaptive Amplifying Factor τ: From Section III-A, we see that the height of the cepstral peak $c_l(q_{F_0})$, which may be seen as an analogue of the signal energy at the harmonics of F_0, is affected by the noise, especially in the higher frequencies. Since speech energy for voiced segments is typically concentrated in the lower frequency regions, an analysis of the harmonic peaks in these regions would provide a better basis for estimating the amplification factor. Therefore, we introduce topographic prominence [14], which provides local information of peaks, for a data-dependent estimate of the scaling factor. A peak is defined as a local maximum and its prominence describes how much the peak stands out from its neighbourhood. The prominence is defined as the vertical distance between the peak and its lowest contour line. Given the excitation spectrum $|\widehat{R}_l(m)|$, the set of local peaks is first identified. Then, a set of prominences $P = \{p_1, p_2, \dots\}$ can be obtained by calculating the prominence for each peak. Fig. 3 shows an example to calculate the ith prominence p_i according to the following steps:

- Extend the peak value horizontally until it crosses the signal or reaches the analysis interval boundary.
- Define the bases of the peak as the lowest values of the signal on each side.
- Define the maximum of the two bases as the contour line.
- The prominence p_i of the ith peak is defined by the vertical difference between the contour line and the peak.

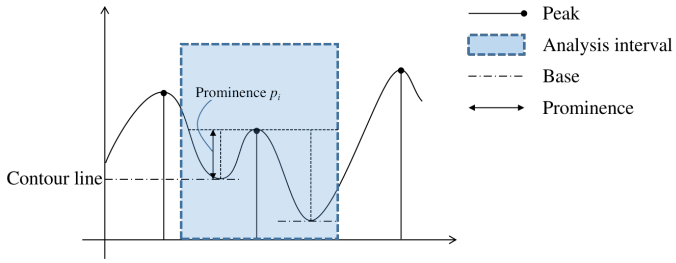


Fig. 3: Topographic prominence: the local peak is extended horizontally to both sides until crossing the signal (the left end) or reaching the pre-determined analysis interval boundary (the right end). Since the left base is higher than the right one, it is used as the contour line to calculate prominence.

Prominence measures the 'local height' of each peak. Since Eq. (12) models the periodic structure by a cosine function, the prominence of the excitation signal is twice the amplitude of this cosine (i.e., the dynamic range of the excitation) in an ideal case. Due to the energy concentration of speech in low frequencies, we assume a high SNR in the first few harmonics and thus true excitation can be recovered after preliminary denoising. The scaling factor τ is then deduced from the prominences whose corresponding peak frequencies are below 1000 Hz:

$$\tau = \frac{\max(p_i | p_i \in P, f_{p_i} \leq 1000\text{Hz})}{2}\,, \quad (13)$$

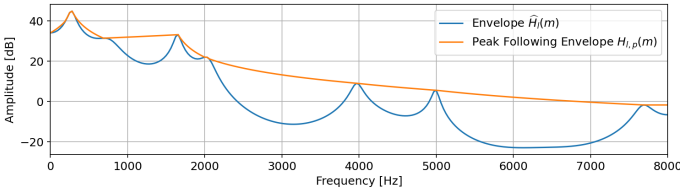
where f_{p_i} is the frequency of the ith peak. The analysis interval of prominence is set to $2 \cdot F_0(l)$ to ensure that the prominence p_i measures the true local height of the harmonics.

We note that in practice noisy fluctuations in the excitation could be recognised as false peaks. However, only peaks at speech harmonics and their vicinity are desired. To ensure this, the minimum distance between two detected peaks is set as $0.8 \cdot M_{F_0}(l)$, where $M_{F_0}(l)$ is the frequency bin index of the fundamental frequency of the current frame. In this distance, only one major peak can be identified.

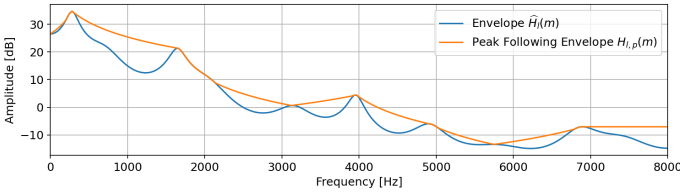
2) Amplitude Decay $\omega(m)$: An extra weighting rule in the frequency domain is introduced to ensure that residual amplitudes are properly tapered down in low-energy bands to generate more natural speech. Intuitively, the dynamic range of harmonics in low-energy bands should be small as well. The speech envelope $|\widehat{H}_l(m)|$ from LPC, which captures speech formants, is a good indicator of the spectral amplitude trend. Therefore, we introduce the weighting rule $\omega(m)$ based on $|\widehat{H}_l(m)|$ to address the mismatch between the synthetic harmonic dynamic range and the speech spectral energy. We first calculate the peak-following envelope $H_{l,p}(m)$ of $|\widehat{H}_l(m)|$ by linear interpolation between two adjacent peaks of $|\widehat{H}_l(m)|$ in the linear domain as shown in Fig. 4. Then, max normalisation

and clipping as shown in Eq. (14) are applied to convert the peak following envelope $H_{l,p}(m)$ into the weighting rule $\omega(m)$. The excitation will be decayed below fundamental frequency after cosine edge decay, so no weighting rule is applied to these frequencies, i.e., $\omega(m) = 1$. When frequency exceeds F_0, the lower boundary of $\omega(m)$ is set as 0.1 to prevent a total harmonic suppression.

$$\omega(m) = \begin{cases} 1, & m \leq M_{F_0}(l) \\ \max\left\{ \dfrac{H_{l,p}(m)}{\max\limits_{m\in\{0,\ldots,M/2\}} H_{l,p}(m)}, 0.1 \right\}, & m > M_{F_0}(l) \end{cases} . \tag{14}$$



(a) An speech active frame, clean



(b) The same speech active frame with pub noise, SNR $= 5$ dB

Fig. 4: Examples of the peak following envelope for the same speech segment under different SNRs.

For the same reason described in section II-C, extra spectrum decay at the excitation spectrum edges is also required in our modification.

Fig. 5 compares enhanced speech spectra by the proposed model (Eq. (12) – Eq. (14)) and the baseline method CEM_{ID}. It can be observed that CEM_{ID} blurs the initial four harmonics due to the insufficient excitation, while the proposed modification is able to suppress the inter-harmonic noise to a satisfactory level. This indicates that the proposed method is able to restore sharpened harmonics even in adverse conditions, which is beneficial to speech quality.

C. Cepstrum Smoothing (CC)

Given the transformation between spectrum and cepstrum in Eq. (5) and Eq. (9), we see that the low-quefrency cepstral coefficients describe the coarse structure (envelope) of the spectrum, while the high-quefrency coefficients include both speech fine structure and noise-related fluctuations. By nulling all the cepstral coefficients in c_{rest}, musical noise is clearly reduced in CEM_{ID}. However, as discussed in [10], this improvement comes at the cost of speech naturalness, because the excitation fine structure is lost in this operation.

Aiming at more naturalness, we propose to improve the CEM_{ID} by preserving more cepstral coefficients. Instead of cepstrum nulling during excitation synthesis, the excitation

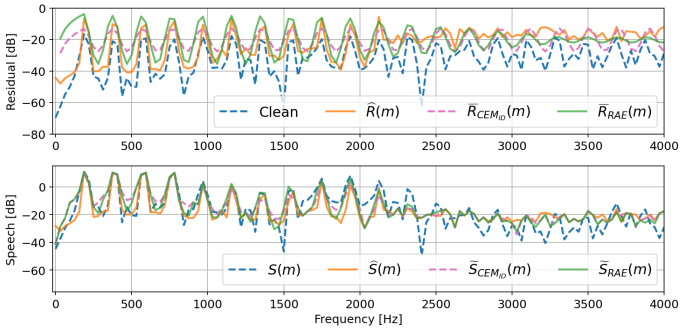


Fig. 5: Demonstration of the benefit of an adaptive pitch amplification factor. The top panel shows the improved speech excitation, and the bottom one shows the benefit of the improved excitation on the final speech estimate for CEM_{ID} and RAE. For comparison, the result of the preliminary denoising and the noisy input frame are also provided. The input is the same voiced frame as in Fig. 2, mixed with white noise at 0 dB. Apart from the clean reference and the preliminary denoising results, the signals enhanced by CEM_{ID} are also provided for comparison. In generating the synthetic excitation, all cesptral coefficients except bin 0 and the bin corresponding to fundamental frequency are discarded. The cepstral value corresponding to the fundamental frequency is scaled up by a constant pitch amplifying factor $\alpha_c = 2$ for CEM_{ID} (from [10]) and by an adaptive factor, computed as proposed in Eq. (12) – (14) for the RAE.

cepstrum is *smoothed* along the quefrency-axis so that more details are preserved while noise is suppressed. We choose to leave coefficients whose quefrencies are lower than a certain threshold quefrency q_{low} unchanged – thereby preserving the general shape – and to average the remaining coefficients with rectangular moving-average windows whose lengths are proportional to the *bit-length* of quefrency bin index q:

$$\bar{c}_{\text{rest},l}(q) = \begin{cases} c_{\text{rest}}(q), & q \leq q_{low} \\ \sum\limits_{i=q-n}^{q+n} \dfrac{n - |q-i| + 1}{n^2} c_{\text{rest}}(q), & q > q_{low} \end{cases} , \tag{15}$$

where $n = \lfloor \log_2(q) \rfloor$. By choosing a window size proportional to the quefrency index, we maintain the averaging interval (for a given sampling rate) even if the DFT analysis window length changes. Rather than define arbitrary functions for window sizes based on the quefrency index, the choice of using the bitlength was an empirical solution that gave the best results based on several trials. We term this method as cepstral convolution (CC).

Fig. 6 compares the proposed cepstral smoothing scheme with CEM_{ID}. It can be observed that our method retains more spectral structure, especially in low frequencies. Another advantage of preserving more cepstral coefficients is to partially compensate for the quantisation error of the pitch detection and the attenuation of the cepstral coefficient corresponding to F_0 in low SNRs.

The two modifications (RAE and CC) enhance speech

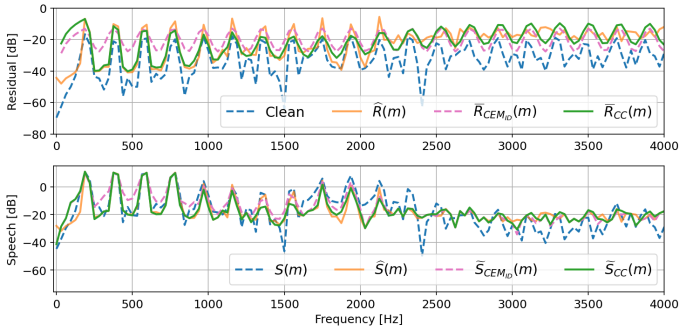


Fig. 6: Demonstration of the benefit of preserving cepstral fine structure. The input is the same voiced frame as in Fig. 2, mixed with white noise at 0 dB SNR. To generate the synthetic excitation with harmonic enhancement, a constant amplification factor of 2 is applied to c_{pitch}. The c_{rest} is set to zeros as proposed for CEM_{ID}; or smoothed by Eq. (15) for the CC approach.

harmonics from complementary aspects. The result of the combined estimation is shown in Fig. 7, where the peaks at the harmonics follow that of the clean speech and the valleys are well accentuated, resulting in a better inter-harmonic noise suppression. Please also refer to Fig. 8 for an appreciation of how these modifications help with the final goal of speech enhancement.

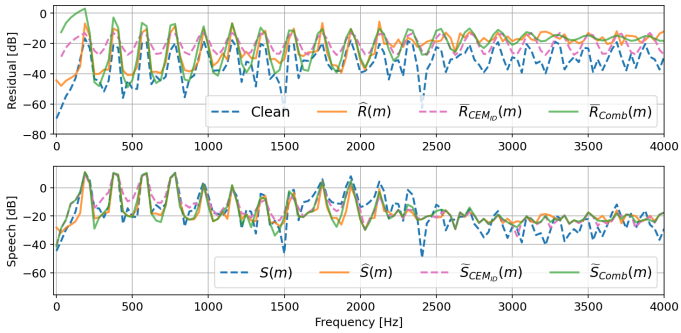


Fig. 7: Demonstration of the benefit of the combined method (applying RAE and CC) to generate the enhanced excitation. The input is the same voiced frame as in Fig. 2, mixed with white noise at 0 dB SNR.

IV. Evaluation

The examples in Fig. 5 – Fig. 7 were chosen to visually demonstrate the benefits of the proposed modifications within the framework of CEM_{ID}. Now we present a more rigorous evaluation of the proposed improvements.

We use the MMSE-LSA gain function with the DD approach as preliminary noise reduction. The four different excitation manipulation methods discussed previously (summarised in Table I) are systematically compared. The baseline approach, CEM_{ID}, is implemented as proposed in [10] with cepstrum nulling and a constant harmonic amplifying factor of $\alpha_c = 2$. The first variant is Residual Amplitude Estimation (RAE) which replaces the constant harmonic amplifying factor by a data-adaptive factor. Our second variant, Cepstral Convolution (CC), still adopts the constant amplifying factor, but smooths cepstral coefficients, i.e., retaining more spectral information, as explained in section III-C. Compared with the baseline CEM_{ID}, the results of these two methods show the respective improvement accrued due to each *individual* modification. Finally, the joint benefit of combining both modifications is evaluated.

TABLE I: Evaluated methods

Method	c_{rest} manipulation	Harmonic synthesis
CEM_{ID}	Eq. (8) (nulling)	Eq. (8) with $\alpha_c = 2$
RAE	Eq. (8) (nulling)	Eq. (12) – (14), adaptive estimation
CC	Eq. (15) (smoothing)	Eq. (8) with $\alpha_c = 2$
Comb	Eq. (15) (smoothing)	Eq. (12) – (14), adaptive estimation

A. Experimental Setup

The four methods were evaluated on the PTDB-TUG database [15]. The database contains clean utterances from 20 speakers (10 males and 10 females). For each speaker, five sentences were randomly chosen from the corpus. Five different signals from the ETSI noise database [16] were mixed with the clean speech at six different SNRs: $\{-5$ dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB$\}$. The noise signals were: *white noise*, *car* (stationary, low-frequency noise), *highway* (non-stationary, low-frequency noise), *buccaneer 1* (narrow-band noise), and *pub* (babble noise). All the speech and noise signals were down-sampled to 16 kHz. To mix the signals at the chosen SNR, the level of the clean speech was measured by the active speech level (according to ITU-T P.56 [17]), and that of noise by the long-term root-mean-square (RMS) value.

For all methods, the frame length is 512 points with a 50% overlap between frames. A square-root von Hann window is used for the analysis and the synthesis. The DFT length is $M = 512$ samples. As the benchmark approach, the parameters of CEM_{ID} are identical to [10]. The order of LPC and the parameter q_{low} for our proposed method CC in Eq. (15) are set to 20 and 10, respectively.

To avoid gain function overflow, the *a posteriori* SNR is limited between -40 dB and 40 dB, and the lower boundary of *a priori* SNR is -25 dB. For both preliminary denoising and final speech estimation, the gain function is limited between -15 dB and 0 dB. The smoothing factor α of the DD approach is 0.98 for preliminary noise reduction.

B. Noise estimation

Instead of using the MS noise estimator as proposed in [10], the Speech Presence Probability Minimum Mean-Square Error (SPP-MMSE) approach with fixed priors [18] is adopted in our work. It has been noted in [18] that MS suffers from noise floor overestimation and delay, while SPP-MMSE is capable of un-biased and fast noise tracking. Note that this noise estimator is used for *all* the evaluated approaches in Table I. As suggested in [18], we assume an equal *a priori* probability for speech presence and absence $P(\mathcal{H}_0) = P(\mathcal{H}_1)$ without prior knowledge, and the optimal *a priori* SPP is set to 15 dB for SPP-MMSE.

C. Quality Measures

First, the same measures as in [10] are employed for a direct comparison, namely noise attenuation, speech-to-speech-distortion ratio, and ΔSNR.

Additionally, noise attenuation and ΔSNR of *the speech active frames* are also employed to highlight the benefits of the excitation manipulation methods. The quality of the different methods is evaluated by the white-box approach [19]. Having obtained the gain function to denoise the observed noisy signal, it is then *separately* applied to the noise component $v(n)$ and to the clean speech component $s(n)$. The measures are subsequently based on the filtered noise $\check{v}(n)$ and the filtered speech $\check{s}(n)$.

The segmental noise attenuation (NA) is calculated as

$$\mathrm{NA}_{\text{seg}} = 10\log_{10}\left[\frac{1}{L}\sum_{l=0}^{L-1}\mathrm{NA}(l)\right], \tag{16}$$

with

$$\mathrm{NA}(l) = \frac{\sum_{k=0}^{T-1} v(k+lT+\Delta)^2}{\sum_{k=0}^{T-1}\check{v}(k+lT)^2}, \tag{17}$$

where T is the frame length, Δ is to compensate the sample delay of the filtered signal (after overlap-add synthesis) and L is the total number of frames in the signal. Higher NA indicates better noise reduction ability of the evaluated method.

The segmental speech-to-speech-distortion ratio (SSDR) measures the distortion introduced by speech enhancement. Higher SSDR suggests less distortion in speech. For frame l, the speech distortion is defined as

$$e(k+lT) = \check{s}(k+lT) - s(k+lT+\Delta),\ k \in (0, T-1). \tag{18}$$

The single frame $\mathrm{SSDR}(l)$ is given by

$$\mathrm{SSDR}(l) = 10\log_{10}\left[\frac{\sum_{k=0}^{T-1} s(k+lT)^2}{\sum_{k=0}^{T-1} e(k+lT)^2}\right]. \tag{19}$$

Since the speech distortion should only be evaluated on the speech active frames, the measured SSDR is the average $\mathrm{SSDR}(l)$ on the set of speech active frames L_1:

$$\mathrm{SSDR} = \frac{1}{||L_1||}\sum_{l\in\{L_1\}}\mathrm{SSDR}(l), \tag{20}$$

where $||L_1||$ is the cardinality of L_1.

In addition, the difference between noisy signal SNR, SNR_{in}, and the denoised signal SNR, $\mathrm{SNR}_{\text{out}}$, provides us with global information about the SNR improvement of the methods. The speech level is calculated according to the active speech level measure from ITU P.56 [17] and the noise level is taken from the long-term RMS value of the noise signal. SNR_{in} is decided by the difference between the active levels of its two components, $s(n)$ and $v(n)$. After the noise reduction, $\mathrm{SNR}_{\text{out}}$ is obtained in the same manner by the two *filtered* components, $\check{s}(n)$ and $\check{v}(n)$.

$$\Delta\mathrm{SNR}(l) = \mathrm{SNR}_{\text{out}}(l) - \mathrm{SNR}_{\text{in}}(l). \tag{21}$$

Further, to investigate the noise reduction ability in speech active frames, NA of speech active frames NA_{act} is calculated on the set of speech active frames L_1:

$$\mathrm{NA}_{\text{act}} = 10\log_{10}\left[\frac{1}{||L_1||}\sum_{l\in\{L_1\}}\mathrm{NA}(l)\right]. \tag{22}$$

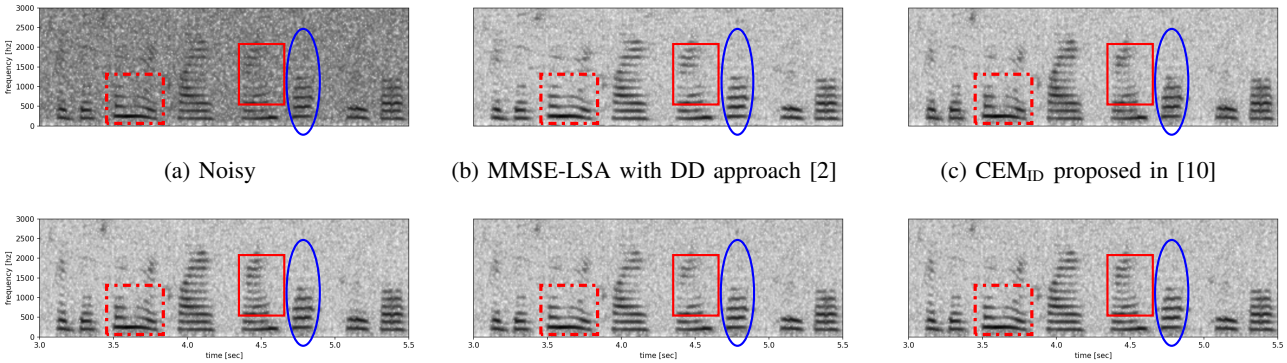
Similarly, the SNR improvement in active frames $\Delta\mathrm{SNR}_{\text{act}}$ is given by

$$\Delta\mathrm{SNR}_{\text{act}} = \frac{1}{||L_1||}\sum_{l\in\{L_1\}}\left[\mathrm{SNR}_{\text{out}}(l) - \mathrm{SNR}_{\text{in}}(l)\right]. \tag{23}$$

The perceptual quality of the filtered clean speech components $\check{s}(n)$ and of the denoised signals $\widetilde{s}(n)$ are evaluated by the wide-band Perceptual Evaluation of Speech Quality (WB-PESQ) [20]. The output of this metric is the mean opinion score - listening quality objective (MOS-LQO). PESQ MOS-LQO scores range from 1.04 to 4.64 and a higher score indicates better speech quality.Note that in the following we drop 'MOS-LQO' in the metric notations for conciseness, but the results are always on MOS-LQO scale rather than raw PESQ scores.

Two metrics, $\mathrm{PESQ}_{\text{st}}$ and ΔPESQ can be derived from the PESQ MOS-LQO score. ΔPESQ is defined as the PESQ score improvement of the enhanced signal compared to the noisy input. It is a comprehensive metric that takes all kinds of artefacts in the processed signal into consideration. $\mathrm{PESQ}_{\text{st}}$ is the score of the filtered speech component. It illustrates the speech distortion introduced by the noise suppression gain function. We can see from the definition that $\mathrm{PESQ}_{\text{st}}$ is unable to detect insufficient noise reduction, e.g., it cannot reflect the gain function overestimation of CEM_{ID} in inter-harmonic frequencies; however, this overestimation indeed leads to a noticeable vocoding effect. Therefore, we additionally introduce $\Delta\mathrm{PESQ}_{\text{act}}$ to compare the audio quality in speech *active* frames. For this, the PESQ scores of the noisy input and the *enhanced* signal $\widetilde{s}(n)$ are evaluated in the speech active frames $l \in L_1$. Before computing the PESQ, all the speech inactive frames ($l \notin L_1$) in the noisy and the enhanced signal are replaced by the silence from the clean utterance. Thus, $\Delta\mathrm{PESQ}_{\text{act}} = \mathrm{PESQ}_{\widetilde{s},\text{act}} - \mathrm{PESQ}_{y,\text{act}}$ reflects the speech quality improvement by the tested methods in the speech active frames.

It should be noted that PESQ was initially designed to measure speech quality degradation in telecommunication [20] and thus it is *not*, in general, a good metric to evaluate speech quality after noise suppression. To have a better idea of the speech quality, we also tested the methods by Perceptual Objective Listening Quality Analysis (POLQA) metric, which is specifically designed for enhanced speech quality evaluation [21]. It allows for predicting speech quality over various distortions for wideband and super-wideband speech signals. Lastly, short-time objective intelligibility (STOI) [22] is employed to evaluate the intelligibility of the denoised signal.



(a) Noisy (b) MMSE-LSA with DD approach [2] (c) CEM_{ID} proposed in [10]

(d) Residual amplitude estimation (RAE) (e) Cepstral convolution (CC) (f) Comb

Fig. 8: Example spectra contrasting the different methods for speech harmonics enhancement in a white noise condition at SNR $= 0$ dB. The spectra on the first row depict the noisy utterance and the results of the two baseline approaches. The spectra on the second row present the results of the proposed methods. **Dash dot square**: this region shows the benefit of preserving more fine structure. **Solid square**: this region highlights the effect of emphasising the fundamental and harmonic frequencies properly. **Oval**: this region best illustrates the drawbacks of a fixed amplification factor and cepstral nulling followed in CEM_{ID}. Detailed discussion can be found in Section IV-D.

D. Experimental Results and Discussion

We start with visual examples for an intuitive appreciation. Fig. 8 shows the spectrograms of the noisy input, MMSE-LSA baseline approach, the CEM_{ID} baseline and the proposed excitation manipulation approaches. This is for the case when a clean speech utterance is mixed with white noise at 0 dB. Generally speaking, all four approaches recover the harmonics to a different degree. It can be observed when comparing Fig. 8d and Fig. 8e with the preliminary denoising result Fig. 8b and the original CEM_{ID} in Fig. 8c that our methods, adaptive harmonic enhancement and cepstral convolution, effectively address the previously discussed weaknesses of CEM_{ID}, and the combination of the two (Fig. 8f) takes the advantages of both proposed modifications to yield an even superior result. For ease of exposition, three regions, where the respective contributions and advantages of the proposed approaches can be appreciated best, are highlighted. **Dash dot square**: this region shows the benefit of preserving more fine structure in the synthesised excitation. Compared to Fig. 8b, the second and the third harmonics in Fig. 8c and Fig. 8d become weaker and discontinuous in time, indicating the drawback of methods purely focusing on the emphasising the F_0 and its harmonics, and neglecting the fine structure. In contrast, preserving this structure by cepstral convolution (Fig. 8e) yields a more time-continuous harmonic structure. The combination of adaptive harmonic enhancement and cepstral convolution (Fig. 8f) yields the best result by combining the advantages of the two manipulations. **Solid square**: this region highlights the effect of *adequately* emphasising the frequencies corresponding to F_0 and the harmonics. Whereas CEM_{ID} is able to boost the low-amplitude harmonics to a certain extent, the benefit of an adaptive amplification factor is evident (Fig. 8d and Fig. 8f). **Oval**: this region best illustrates the drawbacks of a fixed amplification factor and cepstral nulling followed in CEM_{ID}. The insufficient inter-harmonic noise suppression is clearly visible and leads to an audible vocoding effect. RAE shows a slight improvement, while CC and the combined method generate the best results. In the following, the baselines and the proposed modifications are thoroughly evaluated by the metrics introduced in section IV-C. The results are grouped by SNRs of the input signals for detailed performance comparison.

1) Overall Average of Instrumental Metrics: Tables II and III show the measures averaged over the whole test data. Table II shows only the metrics employed in [10], whereas Table III presents the metrics, additionally, on speech active frames.

TABLE II: Evaluation results: instrumental metrics utilised in [10], averaged on the whole test set

method	NA[dB]	SSDR[dB]	ΔSNR[dB]	PESQ_{st}
LSA	11.88	12.23	8.52	3.66
CEM_{ID}	12.20	12.89	8.44	3.67
RAE	11.72	13.59	8.36	3.66
CC	12.04	13.27	8.88	3.59
Comb	11.37	13.86	8.50	3.66

TABLE III: PESQ MOS-LQO and other instrumental metrics on speech active frames, averaged on the whole test set

method	NA_{act}[dB]	$\Delta\text{SNR}_{\text{act}}$[dB]	ΔPESQ	$\Delta\text{PESQ}_{\text{act}}$
LSA	9.12	6.45	0.38	0.54
CEM_{ID}	8.91	5.97	0.39	0.55
RAE	8.58	5.98	0.41	0.60
CC	9.24	6.82	0.41	0.58
Comb	8.66	6.44	0.42	0.61

According to Table II, CEM_{ID} yields higher NA than the other speech enhancement methods, and higher SSDR than the preliminary noise reduction methods. However, a different result is seen when focusing on the speech active frames (Table III). Here CEM_{ID} does not provide benefit (in terms of NA and SSDR) over the baseline MMSE-LSA approach. This divergence indicates that, in terms of NA, CEM_{ID} benefits mostly from the extra noise reduction ability in silent frames

when discarding the majority of the cepstral coefficients of the excitation signal. In terms of SSDR, this indicates the effect of inadequate harmonic emphasis and discarding spectral fine structure.

Comparing with CEM_{ID}, the proposed methods, RAE and CC, introduce less speech distortion, as indicated by a higher SSDR. Cepstral convolution provides good noise reduction in both the global sense and speech active frames. The combined method provides the best speech estimate in terms of speech quality (PESQ MOS-LQO). It shows the lowest NA but similar NA_{act} to CEM_{ID}, which suggests the overall metrics of the combined method are influenced by its performance in silent regions. This trade-off is expected since we also enhanced silent and unvoiced frames and thus generated false harmonics in them.

We described the excitation dynamic range underestimation of CEM_{ID} in section III-A with an example of Fig. 8c; however, it is difficult to observe the degradation of this vocoding effect from the metrics in Table II. The reason is that SSDR and PESQ_{st} are both evaluated on the filtered speech component, so insufficient inter-harmonic gain suppression will not cause artefacts in the filtered speech component (in contrast, it would actually be beneficial for these metrics!) This weakness of CEM_{ID} is, however, reflected by the comprehensive PESQ metrics: ΔPESQ and PESQ_{act}. Compared to the MMSE-LSA baseline, we see only a small benefit from CEM_{ID} in these metrics, whereas the proposed methods yield the best scores.

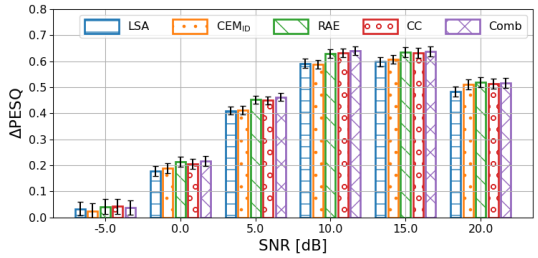
The SNRs of the test set lie in a wide range. To better appreciate the contributions of the various methods we now consider the results grouped by SNR.

2) $\Delta PESQ$ and $\Delta PESQ_{act}$: Fig. 9 provides the results on ΔPESQ and $\Delta\text{PESQ}_{\text{act}}$ at different SNRs (mean improvement and the 95% confidence intervals). We provide the results in this manner because ΔPESQ indicates the overall performance over the noisy utterance (factoring in, thereby, possible degradations introduced by the methods due to errors in F_0 estimation and over-estimation of the harmonic amplitudes), whereas $\Delta\text{PESQ}_{\text{act}}$ is computed on *speech-active* frames. Thereby, $\Delta\text{PESQ}_{\text{act}}$ can better highlight the benefits of the proposed methods on the parts of the signal where they are expected to contribute most prominently.

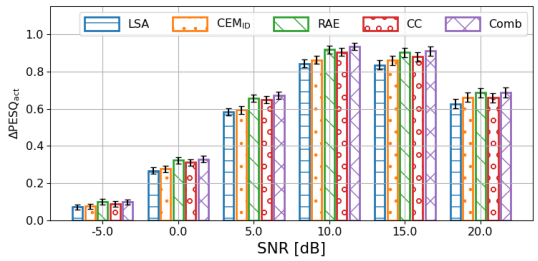
The following insights are obtained. Firstly, in terms of overall quality, we may conclude that while CEM_{ID} yields an average ΔPESQ improvement compared to the baseline MMSE-LSA approach, this is only true for high SNRs (10 dB, 15 dB and 20 dB). At lower SNRs, CEM_{ID} introduces more distortion than this baseline. However, if we consider the confidence intervals, it may be disputed whether this difference is statistically significant. In contrast, each of our proposed improvements consistently provides a better ΔPESQ compared to MMSE-LSA and CEM_{ID}. Especially at SNRs from 5dB to 15dB, the difference may be considered statistically significant.

When we restrict the metric evaluation to the speech active frames ($\Delta\text{PESQ}_{\text{act}}$), CEM_{ID} shows a consistent improvement over the MMSE-LSA baseline for a wider range of input SNRs, but it is again debatable whether this difference is significant (strongly overlapping confidence intervals). In contrast, again, our proposed modifications have higher scores, with the differences being significant over the same SNR range as for ΔPESQ. On the basis of the *PESQ scores*, however, it is difficult to conclude whether the *combination* of RAE and CC (which has the highest score in all conditions) is a significant improvement over the two modifications considered individually. But, this is at least a first indication that the RAE and CC provide *complementary* improvements.

Meanwhile, the improvement in PESQ MOS-LQO of CEM_{ID} gradually approaches that of RAE as SNR increases, which indicates that the proposed RAE is a better candidate for the synthesised excitation, since CEM_{ID} can provide a good estimation under high SNRs.



(a) Average ΔPESQ at different SNRs

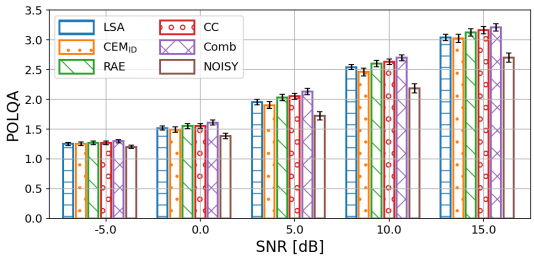


(b) Average $\Delta\text{PESQ}_{\text{act}}$ at different SNRs

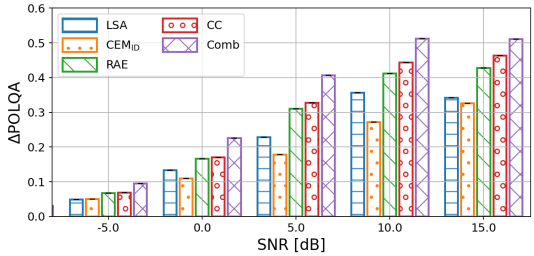
Fig. 9: Improvements in PESQ MOS-LQO on the whole signal (Fig. (a)), and on the speech active frames (PESQ_{act}) (Fig. (b)) for the different methods. The scores are averaged over the different noise types at each SNR. The error bars represent the 95% confidence interval.

3) POLQA: We also evaluated the performance of the five methods using POLQA, which is the new industry standard metric for benchmarking the voice quality for voice communications applications. The evaluation was performed on a subset of 500 gender- and noise-balanced samples with SNRs from -5 dB to 15 dB, which we believe are the most essential SNRs to observe the difference between methods according to the results of PESQ and PESQ_{act}. As shown in Fig. 10, POLQA shows a similar trend as PESQ MOS-LQO: CEM_{ID} degrades speech quality in low SNRs, while the proposed methods are able to improve it under all conditions. In terms of POLQA we see that RAE and CC are consistently better than both baselines: MMSE-LSA and CEM_{ID}, and this performance is significant from an SNR of 5dB onwards. However, RAE and CC, compared to each other, seem to offer the same performance in terms of POLQA. It is now interesting to see that the *combined* method is again better than both RAE and CC, and this difference is significant. This is a strong

indication of the complementary nature of the improvements offered by RAE and CC. The results of POLQA analysis have also been confirmed by a listening test by two experts. A Spearman's correlation coefficient of 0.91 between the expert scores and those of POLQA was found. These results are more reliable indicators of the quality improvements obtained, and reinforce our conclusions based on the other metrics.



(a) Average POLQA at different SNRs

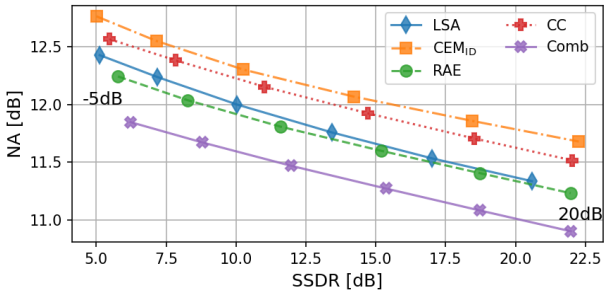


(b) Average ΔPOLQA at different SNRs

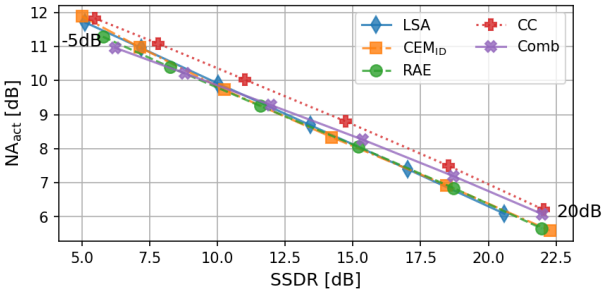
Fig. 10: POLQA and ΔPOLQA of signals enhanced by the different methods, averaged over noise types at each SNR. The proposed methods (RAE, CC and combined methods) are able to improve speech quality in all cases. The combined method shows the highest improvement. The error bars represent the 95% confidence interval.

4) NA and SSDR: Fig. 11 and Fig. 12 demonstrate the effect of the proposed methods on the filtered individual components. Higher NA indicates better noise suppression of the method while higher SSDR indicates less distortion being introduced on the resultant speech during the processing. As expected, noise reduction as well as speech distortion decreases when SNR increases for all methods. It should be noted that in Fig. 11 NA decreases only 1 dB from the lowest SNR cases to the highest SNR cases, while NA_{act} decreases by 5 or 6 dB from one extreme to another. This means a higher overall gain function in the speech active frames, which is expected in speech enhancement.

It is difficult to appreciate the difference among the methods because Fig. 11 shows mainly the magnitude change. Therefore, choosing the preliminary denoising approach (MMSE-LSA) as the baseline, Fig. 12 illustrates the improvement of NA, NA_{act} and SSDR of all excitation manipulation methods compared to this baseline. The performance difference on NA (Fig. 12a) and NA_{act} (Fig. 12b) indicates that CEM_{ID} strongly benefits from extra noise reduction of speech *inactive* frames at low SNRs, as observed from the overall average of these metrics. However, this advantage of NA comes at the cost of extra speech distortion (the lowest SSDR among four methods), which is also more noticeable in low SNR



(a) Average SSDR-NA of each method at different SNRs



(b) Average SSDR-NA_{act} of each method at different SNRs

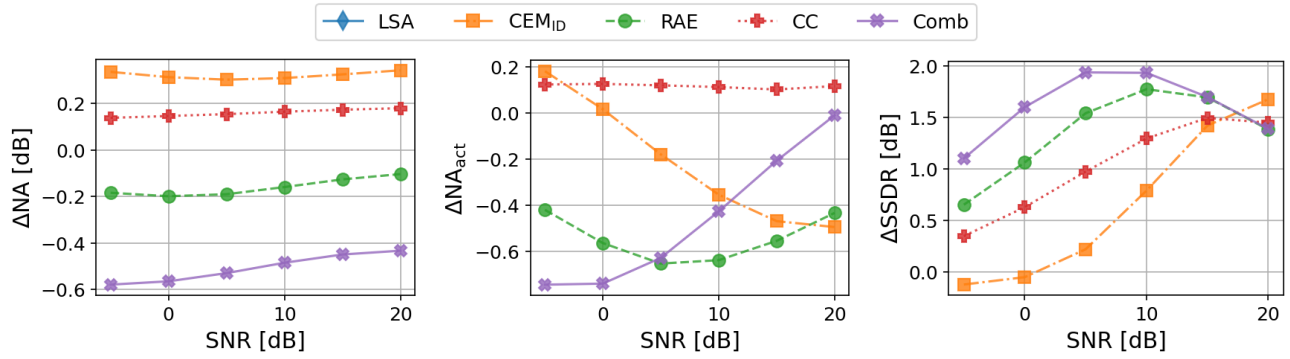
Fig. 11: SSDR-NA relationship of signals enhanced by different methods, averaged among all noise types at each SNR

cases. The combined method scores lower than other methods in noise reduction. Note that CEM_{ID} or our modifications are always carried out without explicit voiced and unvoiced detection. The lower noise reduction score of the combined approach may, therefore, be due to the over-amplification under the harmonic structure assumption in unvoiced and speech inactive frames. In terms of SSDR, the improvement of CEM_{ID} increases sharply as the SNR increases. SSDR of CEM_{ID} at -5 dB is even lower than MMSE-LSA, while our RAE yields more than 1 dB higher SSDR than the LSA baseline in that case. This comparison confirms our hypothesis that using a constant amplifying factor leads to noticeable artefacts due to the underestimation of residual dynamic range. In contrast, both RAE and CC are able to reduce speech distortion. The latter performs better in terms of noise reduction while the former in terms of speech preservation. The combined approach subsequently takes advantage of the complementary nature of these improvements and yields the best result.

5) STOI: The STOI scores, grouped by SNRs, are shown in Fig. 13. The proposed methods exceed CEM_{ID}, but compared to noisy input, only the combined method is able to slightly improve the score in extreme conditions (SNRs of -5 dB, 0 dB, and 5 dB).

V. Conclusions

In this paper we have investigated the CEM approach proposed by [10] in detail. By reformulating the excitation synthesis problem, we were able to get a better insight into the inherent weakness of this approach, namely that the enhanced audio may lose its harmonic sharpness due to dynamic range underestimation and the loss of fine structure



(a) NA (b) NA of the speech active frames (c) SSDR

Fig. 12: Average improvement on NA, NA_{act} and SSDR of methods over preliminary denoising results (by decision-directed approach). The difference between Fig. 12a and Fig. 12b suggests that CEM_{ID} benefits from extra noise reduction in silent regions. Fig. 12c shows that the proposed methods successfully improve speech quality.

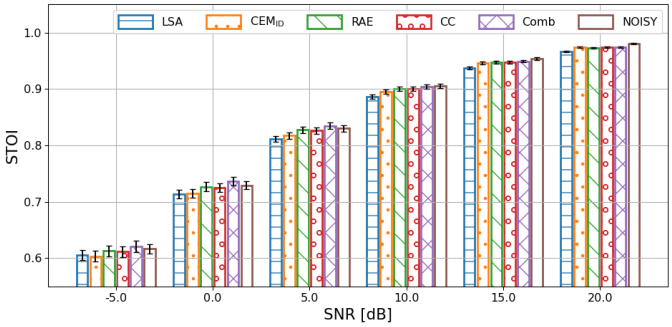


Fig. 13: STOI of noisy signals and enhanced signals, averaged across different noise types at each SNR. The error bars represent the 95% confidence interval.

in the synthesised excitation signal spectrum. Based on our findings, we then proposed two modifications that are able to enhance the harmonic structure of voiced speech in a more natural and robust way. The proposed modifications include residual amplitude estimation and cepstral convolution smoothing. The evaluation results on multi-noise conditions show that the proposed modifications are better able to restore lost harmonics and sharpen the existing ones in voiced frames. Each modification, individually, improves over CEM_{ID}. The two modifications are, also, complementary. This is evident from the fact that the combined method scores higher than each modification individually. The improvement is still robust at low SNRs.

There is still room for further improvement. For example, a finer pitch estimation method could be beneficial. Current F_0 estimation is based on peak-picking on the discrete quefrency bins and then calculating the corresponding frequencies. This could introduce a quantisation error when the actual fundamental frequency falls between the adjacent bins. Secondly, since we assume a harmonic structure for all frames, there is the risk of stronger musical noise, which has been reflected by the difference between metrics on active frames and these on the whole signal. This can be solved by introducing a voice activity detection module, and applying the proposed method to voiced speech frames only.

Lastly, we note that CEM need not be seen as a stand-alone method. In our work, we consider a statistical noise suppression framework within which CEM is integrated. However, in practice, CEM can be piggy-backed onto any denoising framework which can output an estimate of the gain function and noise floor. This also opens the possibility to integrate CEM within DNN-based frameworks, allowing for a marriage of model-based approaches and data-driven approaches, with all the ensuing benefits thereof. These are directions we will consider for the future.

We urge the reader to listen to the audio examples at https://yanjuesong.github.io/Improved-CEM-samples/.

ACKNOWLEDGMENTS

The authors would like to thank the experts N.M.P. Neumann and J.G. Beerends at TNO Netherlands for helping with the POLQA scores and for conducting the listening tests.

REFERENCES

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

[2] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[3] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2098–2108, 2006.

[4] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral smoothing of spectral filter gains for speech enhancement without musical noise," *IEEE Signal processing letters*, vol. 14, no. 12, pp. 1036–1039, 2007.

[5] ——, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4897–4900.

[6] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*. John Wiley & Sons, 2006.

[7] T. Rosenkranz, "Modeling the temporal evolution of LPC parameters for codebook-based speech enhancement," in *2009 Proceedings of 6th International Symposium on Image and Signal Processing and Analysis*. IEEE, 2009, pp. 455–460.

[8] R. Chen, C.-F. Chan, and H. C. So, "Model-based speech enhancement with improved spectral envelope estimation via dynamics tracking," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1324–1336, 2011.
[9] T. Mellahi and R. Hamdi, "LPC-based formant enhancement method in Kalman filtering for speech enhancement," *AEU-International Journal of Electronics and Communications*, vol. 69, no. 2, pp. 545–554, 2015.
[10] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "Instantaneous a priori SNR estimation by cepstral excitation manipulation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1592–1605, 2017.
[11] S. Elshamy and T. Fingscheidt, "DNN-based cepstral excitation manipulation for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1803–1814, 2019.
[12] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain based single-microphone noise reduction for speech enhancement*, ser. Synthesis Lect. Speech and Audio Proc. Morgan & Claypool, 2013.
[13] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on speech and audio processing*, vol. 9, no. 5, pp. 504–512, 2001.
[14] B. Kashyap, M. Horne, P. N. Pathirana, L. Power, and D. Szmulewicz, "Automated topographic prominence based quantitative assessment of speech timing in cerebellar ataxia," *Biomedical Signal Processing and Control*, vol. 57, p. 101759, 2020.
[15] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.
[16] ETSI, "Etsi noise database," https://docbox.etsi.org/stq/Open/EG%20202%20396-1%20Background%20noise%20database/Binaural_Signals. (Last accessed 04/2022), 2006.
[17] ITU-T, *Rec. P.56: Objective Measurement of Active Speech Level*. International Telecommunication Union-Telecommunication Standardisation Sector, 2011.
[18] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2011.
[19] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-optimized speech enhancement," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 4, pp. 825–834, 2008.
[20] ITU-T, *Rec. P.862.2: Corrigendum 1, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*. International Telecommunication Union-Telecommunication Standardisation Sector, 2017.
[21] ——, *Rec. P.863: Perceptual objective listening quality prediction (POLQA)*. International Telecommunication Union-Telecommunication Standardisation Sector, 2018.
[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4214–4217.