EXCITING ESTIMATED CLEAN SPECTRA FOR SPEECH RESYNTHESIS

Sreyas Srimath Tirumala

The Ohio State University
Computer Science & Engineering
Columbus, OH 43210
srimath-tirumala.1@osu.edu

Michael I Mandel

Brooklyn College, CUNY
Computer & Information Science
Brooklyn, NY 11210
mim@sci.brooklyn.cuny.edu

ABSTRACT

Spectral masking techniques are prevalent for noise suppression but they damage speech in regions of the spectrum where both noise and speech are present. This paper instead utilizes a recently introduced analysis-by-synthesis technique to estimate the spectral envelope of the speech at all frequencies, and adds to it a model of the speech excitation necessary to fully resynthesize a clean speech signal. Such a resynthesis should have little noise and high quality compared to mask-based approaches. We compare several different excitation signals on the Aurora4 corpus, including those derived from the high quefrency components of the noisy mixture and from the combination of a noise robust pitch tracker and a voiced/unvoiced classifier. Preliminary subjective evaluations suggest that the speech synthesized using our approach has higher voice quality and noise suppression than spectral masking.

Index Terms— Analysis-by-synthesis; Speech synthesis; Noise suppression; Noise robustness.

1. INTRODUCTION

Spectral masking is a widely used technique for noise suppression. The approach aims to reduce noise by attenuating the time-frequency points in the noisy signal's spectrogram that are dominated by noise [1]. The Ideal Binary Mask (IBM) has been proposed as the goal of computational auditory scene analysis [2, 3]. It is computed from the clean speech and noise signals before they are mixed together by assigning a gain of 1 to all time-frequency points with SNR above a certain threshold, and 0 to all others. While such a mask cannot be computed for real noisy recordings, many systems exist to estimate such masks for them [4–6, e.g.,]. Even in the ideal case, these methods suffer from two problems with respect to clean speech re-synthesis. First, speech is attenuated in regions where there is loud noise, but still significant speech energy. Second, since speech alternates between periods of activity and inactivity in each frequency band, noise will be completely attenuated in periods of speech inactivity, but incompletely attenuated in periods of speech activity, leading to a disturbing noise pulsation.

The human speech production system is generally represented by the source-filter model [7]. In this model, the speech source comes from the vibration of the vocal folds, which excite the resonances of the vocal tract filter. We handle the clean speech synthesis process similarly as a two part problem, first of obtaining a noise robust estimate of the spectral envelope and second of generating an adequate or realistic excitation [8]. If both the spectral envelope and the excitation can be estimated accurately from the noisy signal, then the synthesized signal should be comparable to the original clean speech without any noise or distortion.

In this paper we use a recently introduced analysis-by-synthesis technique [9] to generate an estimate of the mel frequency cepstral coefficients (MFCC) of clean speech from the noisy speech signal. Given a spectral mask, it estimates the clean MFCC's by minimizing a combination of the distance between the estimated spectrum and the masked noisy observation, and the negative log likelihood under a large vocabulary continuous speech recognizer (LVCSR). This technique regularizes the spectral estimation with the likelihood under the LVCSR, thereby forcing the estimated MFCC's to be more "speech like". The results in [9] show that this technique estimates spectral envelopes that are closer in Itakura-Saito divergence to the clean spectral envelope than an envelope based on MFCCs directly computed from the masked speech is. Thus these estimated MFCCs are more suitable for clean speech re-synthesis, although to do so requires additionally estimating an appropriate excitation.

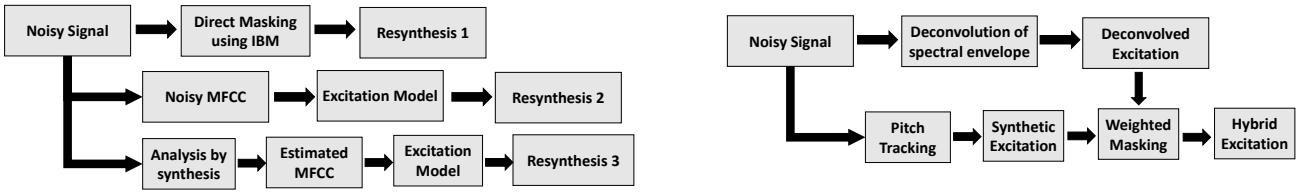
This paper compares three different excitations for this re-synthesis using the estimated MFCCs [10]. The cepstrum of the noisy signal contains a noise-corrupted excitation in the higher quefrency regions. The first excitation comes from the deconvolution of the noisy speech, and the second from the deconvolution of the noisy speech with the spectral mask applied to it. The third excitation which we call the hybrid excitation, combines this direct-masked noisy excitation with a synthesized excitation in regions where the mask is 0 (i.e., where there is no excitation energy in the direct-masked noisy signal). This synthetic excitation uses a pulse train controlled by a noise-robust pitch tracking algorithm [11] in voiced regions and white noise in the unvoiced regions. When combined with the excitation from the speech-dominant masked regions, this results in an signal which not only contains the true pitch with a realistic phase, but also is noise robust and avoids introducing artifacts. Our experiments showed that this hybrid excitation resulted in the highest quality speech signal as compared to the other excitation models.[1]

2. SYNTHESIS USING ESTIMATED SPECTRA

Figure 1(a) shows the resynthesis approaches compared in this paper as a flowchart. Figure 1(b) shows the various excitation signal generation processes.

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1409431. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

[1] Sample results can be found at: http://mr-pc.org/work/waspaa15b



(a) Resynthesis methods

(b) Excitations

Figure 1: Flowcharts showing the computation of the (a) complete resynthesis methods and (b) excitation signals that are combined into the final resyntheses and their inter-relation.

2.1. The Speech synthesis process

Speech production, according to the source-filter model, can be viewed as the result of passing the glottal excitation source through the vocal tract filter. From this model, speech is mathematically described as the convolution of the spectral envelope with a pulse train-like excitation. In this description, the spectral envelope which contains most of the phonetic or linguistic information, is characterized by spectral features such as Perceptual Linear Prediction (PLP) coefficients, Linear Prediction Coefficients (LPC), or Mel Frequency Cepstral Coefficients (MFCC) [7, 12, 13]. These features are widely used for speech recognition and are thus sufficient for identifying this information. They aim to be robust to variations beyond the spectral envelope, and so by design eliminate the excitation information. Because of this design, in order to reconstruct the signal from the MFCCs, the excitation signal has to be generated separately. In order to create a clean-sounding signal, this excitation must appropriately excite the spectral envelope over all of the formants while including consistent phase information across frequencies [13].

2.2. Estimating clean spectral envelope: Analysis-by-synthesis

The recently introduced analysis-by-synthesis approach of [9] poses the estimation of the clean MFCCs as an optimization problem given a spectral mask. The estimated MFCCs are optimized to fit the noisy observation where the mask's gain is high while being regularized to have a high likelihood under an LVCSR, which thereby maintains the speech-like characteristics of the estimated sequence of MFCCs. The sequence of MFCC vectors for the entire utterance, x, is found that optimizes

$$\hat{x} = \underset{x}{\operatorname{argmin}}\, \mathcal{L}(x; M) \tag{1}$$

$$= \underset{x}{\operatorname{argmin}} (1-\alpha)\mathcal{L}_I(x; M) + \alpha\mathcal{L}_H(y(x)) \tag{2}$$

where $\mathcal{L}_I(x; M)$ measures the divergence of the estimated spectra from the noisy observation in regions where the mask, M, is non-zero, $\mathcal{L}_H(y(x))$ is the log likelihood of LVCSR features, $y(x)$ derived from x, under a recognizer trained on clean speech, and α controls the relative importance of the two criteria. Empirically, these estimated MFCCs provide a spectral envelope that is close to the true clean speech envelope.

2.3. Excitation Models

We compare excitation estimates derived from the high quefrency components, quasi periodic pulse trains derived from the pitch information, and a hybrid combination of the two. In the following subsections we will describe all of them briefly. The pitch-based excitation utilizes a noise robust pitch tracking algorithm, which predicts both the pitch and pitch salience function from the observed correlogram [14, 15]. It enforces continuous pitch contours using a hidden Markov model (HMM). Another noise robust pitch tracking method is described in [11] where a neural network is trained on the full auto-correlation pattern of multiple sub-bands coming out of the filterbank and this neural network detects the peaks which indicate particular pitches. Both of these methods are shown to be effective in pitch tracking for noisy speech signals. A quasi periodic pulse train can then be derived from the pitch track of the noisy signal. It should be noted that the unvoiced regions, found using a voiced/unvoiced detector [11], are not considered for the generation of the quasi-periodic pulse train but are excited using white noise and cross-faded with the pulse train. The excitation models and their interrelations are shown broadly in Figure 1(b).

2.3.1. Excitation from deconvolution

The speech signal as mentioned before is a convolution of the source excitation and the vocal tract filter. Deconvolution thus seems a reasonable approach to extracting the excitation. Since the vocal tract filter impulse response is typically much shorter than the STFT analysis window, this deconvolution can be implemented as a simple division of the observed spectrogram by a comparable representation of the spectral envelope. The presence of noise in the observation complicates this procedure, however. We compare two versions of this excitation. The first is created from ratio of the full noisy signal with the spectral envelope estimated by the analysis-by-synthesis system. This excitation contains residual noise, so the second excitation is created by applying the estimated mask used in the analysis-by-synthesis procedure to this first excitation. This eliminates the noise, but leaves holes in the spectrum where the noise is louder than the speech, a problem that will be addressed in Section 2.3.3.

2.3.2. Synthetic excitation from pitch

The time-varying pitch of the speech signal can be used to generate a quasi-periodic pulse train to excite the estimated spectral envelope [16]. We utilize frequency modulation to synthesize a pulse train that smoothly varies in frequency while avoiding instantaneous jumps in phase that could cause audible glitches. The phase information can be calculated from the pitch by taking an integral of the

interpolated pitch over each sample.

$$\phi(t) = \sum_{t'=1}^{t} I[f_0(k)](t') \quad (3)$$

Where t is time, k indexes STFT frames, I is the interpolation function which estimates pitch for each sample and $f_0(k)$ is the frame wise pitch track. From the phase, the pulse train can be generated by:

$$E[n] = \mu(\exp(jT_0\phi(t))) \quad (4)$$

Where μ is a function that is 1 when the rotating sinusoidal phasor sampled at a period of T_0 passes 0 phase, and 0 otherwise. At boundaries, this signal is cross-faded with the white noise excitation used in the unvoiced regions.

2.3.3. Hybrid excitation

Both the above described excitations have shortcomings, the deconvolution excitation suffers from noise leakage in speech-dominated regions whereas the synthetic excitation is unable to capture the finer details of the harmonics of the true vocal fold excitation. Thus we generated a hybrid excitation signal by interpolating them in the time-frequency domain. In regions where the estimated mask is 1, we use the observed excitation, and in regions where it is 0, we use the synthetic excitation to help recover from speech attenuation. The relation between this and the other two excitation models is broadly shown in Figure 1(b).

3. EXPERIMENTS

We evaluated the subjective quality of several different speech resynthesis methods using a MUltiple Stimuli, Hidden Reference and Anchor (MUSHRA) listening test. Systems were compared in terms of subjective speech quality, noise suppression, and overall signal quality.

The noisy speech used in this experiment came from the Aurora4 dataset [17] which contains clean speech of different utterances of the wall street journal medium vocabulary database mixed at 5-15 dB signal-to-noise ratio with six different noise types: car, airport, street, babble, train and restaurant. The Aurora4 dataset contains the original clean signal which we used as a reference. The analysis-by-synthesis system was trained on the clean-only portion of the Aurora4 training set. The listening test was performed on a small subset of the Aurora4 test set.

3.1. Comparison systems

For spectral envelopes, we used the estimated clean MFCCs generated by analysis-by-synthesis and the MFCCs derived directly from the noisy signal as a reference point for comparison. For the excitations, we compared several signals derived from the clean (as an oracle), noisy, and direct-masked signals. Throughout these experiments, we use the spectral masks predicted by [5]. Thus in our experimentation framework we compare the combination of spectral envelopes from estimated clean MFCCs, noisy MFCCs, and MFCCs from the direct-masked noisy speech with excitations generated from clean, noisy and direct masked noisy speech.

We used four different excitation models in our experiments in combination with the two envelopes. These were the deconvolved excitation from the original clean speech signal as a baseline, the deconvolved excitation from the noisy signal, the deconvolved excitation from the direct masked noisy signal, and the hybrid excitation merging the synthetic pitch-based excitation with the deconvolved excitation from the direct-masked signal. The pitch track of the noisy signal was calculated using sub-band auto correlation and the voiced and unvoiced regions of speech were determined from the pitch track calculated on the noisy speech signal [11]. From the pitch track and voiced / unvoiced regions, a synthetic quasi periodic pulse train was generated. The deconvolved excitation was estimated on both the clean and noisy signals by dividing their spectrograms by the estimated spectral envelope derived from their MFCCs.

3.2. Listening tests

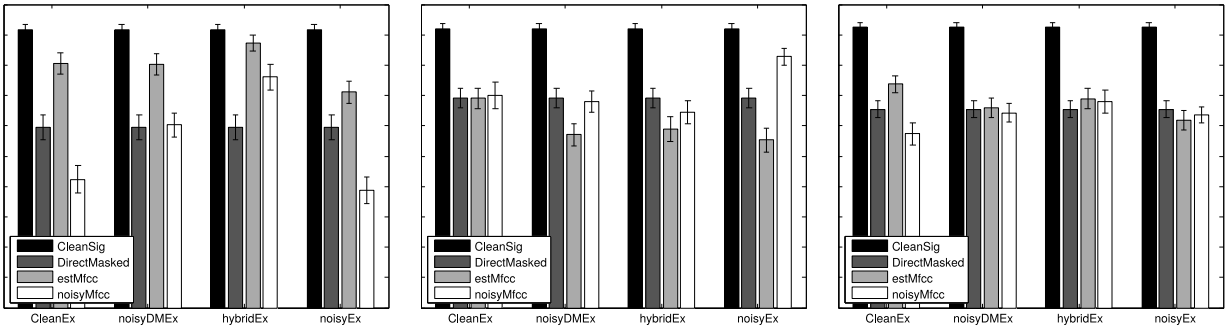
We performed listening tests to compare our system with the noisy MFCCs and the direct masking approach. Initial analysis of the Short Term Objective Intelligibility metric (STOI) [18] showed that our method and direct-masking showed very high ($>$0.95) predicted intelligibility but listening to the signals suggested differences in noise suppression, hence we performed more formal listening tests to measure the speech quality and noise suppression. The data used for the listening tests contained the original clean speech signal, the noisy signal, the direct masked noisy signal and the re-synthesized signals using the estimated MFCCs, noisy MFCCs and four different excitation models. In total 15 versions of each of 30 utterances (5 different utterances mixed with 6 types of noise each) were used for the listening test. The signals were selected in no particular order from the testing set before listening to them.

We used the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) paradigm for the speech quality tests. For each of the 30 original utterances, listeners were provided with the clean and the original noisy utterance labeled as such, which act as the anchor signals, and the remaining re-synthesized utterances unlabeled and in a random order. Listeners were asked to asses the speech quality, noise suppression quality, and overall quality of each signal on a scale from 0 (poor) to 100 (excellent). This was contextualized for them by asking them to imagine their reaction to hearing this signal from their stereo or as part of a radio broadcast. The order of the utterances was randomized for each listener and the order of the systems was randomized for each utterance and each listener. A total of six listeners participated in this listening test, four of whom had never heard the Aurora4 dataset before and none of whom had heard the processed files before.

3.3. Results

The initial objective intelligibility analysis using STOI [18] gave very high ratings to both the spectral masked signals and those re-synthesized using our approach, providing support to our intuition that in both cases, the intelligibility of the speech was not affected greatly. Figure 2 shows all the subjective results from the MUSHRA test, averaged over all utterances and listeners. Figure 2(a) shows listeners' rating of the signals in terms of noise suppression quality, Figure 2(b) shows their rating of the quality of the speech, and Figure 2(c) shows their overall rating of the signal taking into account both the noise suppression and speech quality.

Figure 2(a) shows that in terms of noise suppression, the re-synthesized signals using estimated MFCCs greatly outperform those using noisy MFCCs and those from direct masking. This is more strongly supported by the fact that this trend is observed across excitation models. Figure 2(b) shows that the best speech quality comes



(a) Noise suppression (b) Speech quality (c) Overall quality

Figure 2: Results of MUSHRA listening test judging quality of synthesized speech. The cleanEx is the deconvolved excitation from the clean reference signal; noisyDMEx is the excitation by masking and deconvolving the noisy signal; hybridEx is the combination of the masked and synthetic excitation; noisyEx is the deconvolved excitation from the noisy signal. Note that the DirectMasked result does not depend on excitation, but is replicated for comparison purposes.

from the noisy MFCCs and noisy excitation, essentially recreating the noisy signal. This makes sense, as that signal also has the lowest noise suppression quality. The estimated MFCCs sacrifice some speech quality in order to achieve their noise suppression. This is not the case when using the oracle clean excitation, showing that a better estimate of the clean excitation could improve both metrics together. The hybrid excitation produces slightly better speech quality with the analysis-by-synthesis MFCCs than the other excitations do. Figure 2(c) compares the systems on overall quality. It shows that our hybrid excitation model slightly outperforms the rest. While it is inferior in quality in comparison to the clean speech reference, it is superior to the noisy signal, the signal re-synthesized using the estimated MFCCs and the masked excitation, and the direct masked signal.

The above results show that the estimated MFCCs performed better than the rest in terms of noise suppression as evidenced by their high scores even when the deconvolved excitation from the noisy signal was used. Secondly, the estimated MFCCs outperform the other envelopes in terms of overall quality, although not by a lot. Thirdly, the hybrid excitation model outperforms all other excitation models both in terms of quality and noise suppression for both the estimated and noisy MFCCs.

The significance of these results was analyzed using a paired two-sample t-test one between the estimated MFCCs with the hybrid excitation and the noisy signal, and another between the estimated MFCCs with the hybrid excitation and the direct-masked signal. The tests showed that in both cases, the estimated MFCCs with hybrid excitation had significantly higher noise suppression quality ($p < 10^{-6}$ for both), significantly lower speech quality ($p < 10^{-6}$ for both), but significantly higher overall quality than the comparison systems ($p = 0.0004$ vs noisy, $p = 0.024$ vs direct-masked). These results suggest that the estimated MFCCs with hybrid excitation improves noise suppression more than it damages speech quality as they are combined by listeners into overall quality, i.e., it is possible to use the analysis-by-synthesis system of [9] to improve signal quality when combined with this hybrid excitation.

4. CONCLUSION

This paper introduces a new approach for the re-synthesis of clean speech signals from noisy audio by using estimated clean spectral envelope derived from the analysis-by-synthesis approach with various excitation signals. The results of our experiments show that our approach outperforms the standard direct masking approach on the Aurora4 dataset on both noise suppression and overall signal quality. These initial results are promising but, further improvements could be made by generating better estimates of the clean excitation signal, as the results show that overall quality is heavily dependent on the excitation, and the oracle clean excitation leads to excellent quality. Further improvements could be made by jointly estimating both the excitation and the spectral envelope in the analysis-by-synthesis procedure.

5. REFERENCES

[1] A. Narayanan and D. Wang, "On the role of binary mask pattern in automatic speech recognition," in *Proceedings of Interspeech*, 2012.

[2] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, 2005, pp. 181–197.

[3] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Isolating the energetic component of speech-on- speech masking with ideal time-frequency segregation," vol. 120, pp. 4007–4018, 2006.

[4] A. Narayanan and D. L.Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," in *Tech. Rep. OSU-CISRC-6/13-TR14, Ohio State University Department of Computer Science and Engineering*, 2013.

[5] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference*, 2013, pp. 7092–7096.

[6] M. I. Mandel, S. Bressler, B. Shinn-Cunningham, and D. P. Ellis, "Evaluating source separation algorithms with reverberant speech," vol. 18, pp. 1872–1883, 2010.

[7] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, 2nd ed. Prentice Hall, 1993.

[8] M. Cooke, "A glimpsing model of speech perception in noise." vol. 119, pp. 1562–1573, 2006.

[9] M. I. Mandel and A. Narayanan, "Analysis-by-synthesis feature estimation for robust automatic speech recognition using spectral masks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[10] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," pp. 970–990, 1980.

[11] B. S. Lee and D. P. W. Ellis., "Noise robust pitch tracking by subband autocorrelation classification," in *Proc. Interspeech*, 2012, pp. 707–710.

[12] D. P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005. [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

[13] D. Talkin, W. B. Kleijn, and K. K. Paliwa, "A robust algorithm for pitch tracking(rapt)," in *Speech Coding and Synthesis, The Eds. Ams-terdam, Netherlands:Elsevier*, 1995, pp. 495–518.

[14] Z. Jin and D. Wang., "A multipitch tracking algorithm for noisy and reverberant speech," in *Proc. IEEE ICASSP*, vol. 19, 2005, pp. 4218–4221.

[15] M. Wu, D. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," vol. 11, pp. 229–241, 2003.

[16] Z. Ou and Y. Zhang, "Probabilistic acoustic tube : a probabilistic generative model of speech for speech analysis / synthesis," in *Proceedings of International Conference on Artificial Intelligence and Statistics*, 2012, pp. 841–849.

[17] N. Parihar, J. Picone, D. Pearce, and H. G. Hirsch, "Performance analysis of the aurora large vocabulary baseline system," in *Proceedings of EuSIPCo*, 2004.

[18] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of timefrequency weighted noisy speech," vol. 19, pp. 2125–2136, 2011.