

A NOVEL APPROACH TO SOFT-MASK ESTIMATION AND LOG-SPECTRAL ENHANCEMENT FOR ROBUST SPEECH RECOGNITION

Julien van Hout, Abeer Alwan

Electrical Engineering Department, University of California, Los Angeles.
julienvanhout@ucla.edu, alwan@ee.ucla.edu

ABSTRACT

This paper describes a technique for enhancing the Mel-filtered log spectra of noisy speech, with application to noise robust speech recognition. We first compute an SNR-based soft-decision mask in the Mel-spectral domain as an indicator of speech presence. Then, we exploit the known time-frequency correlation of speech by treating this mask as an image, and performing median filtering and blurring to remove the outliers and to smooth the decision regions. This mask constitutes a set of multiplicative coefficients (ranging in $[0,1]$) that are used to discard the unreliable parts of the Mel-filtered log-spectrum of noisy speech. Finally, we apply Log-Spectral Flooring [1] on the lifted spectra of both clean and noisy speech so as to match their respective dynamic ranges and to emphasize the information in the spectral peaks. The noisy MFCCs computed on these modified log-spectra show an increased similarity with their corresponding clean MFCCs. Evaluation on the Aurora-2 corpus shows that the proposed approach competes with state-of-the-art front-ends, like ETSI-AFE, MVA or PNCC.

Index Terms— Speech Recognition, Feature Extraction, Speech Enhancement, Mask Estimation, Median Filtering.

1. INTRODUCTION

Traditional features like MFCCs or LPCCs along with Hidden Markov Model-based statistical engines perform well in speech recognition tasks as long as the training and testing sets are recorded in similar conditions. Yet, because of the increasing need to use recognition engines on mobile devices in different environments, testing sets no longer match the recording conditions of the training data.

The main goal of a noise-robust front-end for ASR is to develop features that retain useful variability in speech while minimizing variability due to the corrupting noise. To this end, interesting work has been carried out on the computation of Speech Presence Probabilities (SPPs) with applications to speech enhancement. Reliable SPPs provide clues about the spectro-temporal location of speech and are thus a highly valuable tool for noise reduction algorithms. Ephraim et al. ([2]-[3]) first proposed a framework for an MMSE-based speech enhancement algorithm relying on SPPs. Recent work on estimating SPPs relies on statistical modeling of the speech and noise signals. Attempts have been made to exploit the well-known time-frequency correlation of speech signals by smoothing the SPPs with an HMM [4], [5]. Some other notable work has been done in the realm of soft mask estimation, a problem closely related to SPP estimation, where the coefficients are also within $[0,1]$, but do not represent actual probabilities. A hybrid approach to soft mask computation is proposed in [6], which labels

unvoiced frames using a spectral subtraction-based mask, and labels voiced frames by extracting information from the harmonics.

Once these speech presence indicators are available, either in the form of actual SPPs, soft masks, or binary masks, one needs to exploit this information to enhance the signal. To this end, other work has been done in the domain of data imputation. The latter framework makes use of either a soft mask or a binary mask that labels time-frequency bins as more or less reliable. Then, the information about the unreliable bins is mostly erased, and new values are filled in using information about the reliable parts of the spectrum as well as some side knowledge about the underlying speech signal. Among the most successful approaches, [7] proposes to model prior-knowledge about the speech with GMM-based distributions, before inferring values for the unreliable data. A very recent investigation [8] showed that masking the spectrum with an ideal binary mask leads to better results than [7] on tasks requiring a strong language model, with a much-reduced computational cost. Finally, recent approaches feature the use of compressive sensing as a means for data recovery under the assumption of sparsity of the clean speech signal. In [9] and [10], a basis for sparsity is obtained by accumulating a large dictionary of exemplars whereas [11] and [1] exploit the time-frequency correlation of speech and use an image processing inspired two-dimensional Haar transform on the spectrographic data.

In this paper, a framework for soft mask based speech enhancement partly inspired from previous studies is proposed. We exploit the spectro-temporal correlation of speech to perform accurate reconstruction of the signal at a low computational cost. This paper makes the two following contributions. First, we derive a simple SNR-based soft mask that we further enhance using basic image processing techniques such as median filtering and blurring. This approach allows us to remove the SNR outliers due to noise variability and helps localize weak but spatially coherent speech areas even in adverse noise conditions. The resulting values are not meant to be actual probabilities, but smoothed indicators of the relative strength of speech and noise at each time-frequency bin. The second contribution lies in the use of soft decision masks as a computationally inexpensive way of performing enhancement. We show that using our mask as a set of multiplicative weights on the log-spectrum efficiently discards the noise while retaining most of the speech information. A complementary flooring step [1] is performed to match the respective dynamic ranges of both the original clean and the enhanced noisy spectra.

This paper is organized as follows: The SNR-based soft decision mask is presented in Section 2, the enhancement procedure in Section 3. Section 4 provides experimental results. Finally, conclusions are presented in Section 5.

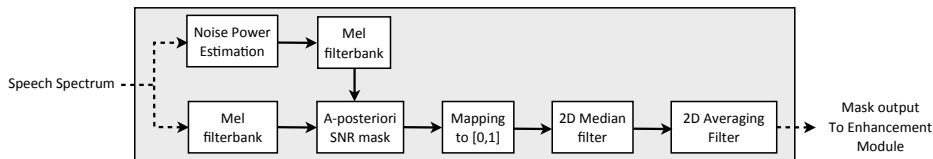


Fig. 1. Flowchart of Soft-Mask computation

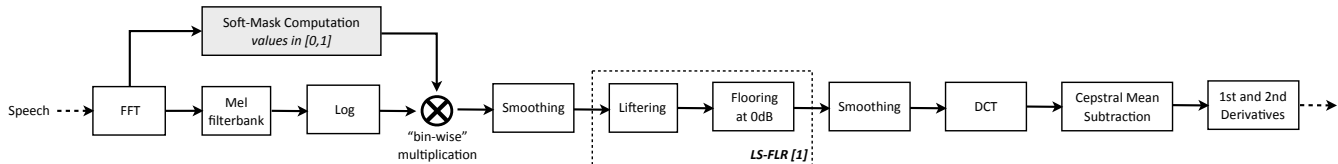


Fig. 2. Flowchart of the complete enhancement-based front-end

2. SOFT-DECISION MASK

In this section, we describe the steps in the computation of the SNR-based soft decision mask. The mask comprises values between 0 and 1, where a value close to 1 indicates that the energy in the noisy signal originates from speech in a significantly high proportion. A flowchart is shown in Fig.1.

2.1. From channel-wise SNR to a first soft-mask estimation

We first perform an estimation of the noise power in each of the 32 Mel-channels, by averaging the energy over the first and last 15 frames. Then, we compute an estimate of the a-posteriori SNR γ at each time-frequency bin.

$$\gamma = 10 \cdot \log \left[\max \left(\rho_{\min}, \frac{P_{\text{signal}}}{P_{\text{noise}}} \right) \right]$$

where P_{signal} and P_{noise} are the respective powers of the signal and the noise. The threshold $\rho_{\min} = 0.5$ serves as a flooring value.

The above estimation relies on the assumption that the noise is stationary and on the oracle information that the first and last 15 frames are composed solely of noise. Yet, one might object to the validity of these two assumptions, especially in environments where the noise is often rising or fading and where its power should regularly be re-estimated. For the latter case, Section 2.3. describes a solution based on [12], which can be incorporated into our framework in order to handle time-varying noise conditions.

The SNR estimate is mapped to the interval [0,1] by using the tunable sigmoid function

$$f(x) = \frac{1}{1 + \exp^{-\alpha(x-\beta)}} \quad (1)$$

where $\alpha = 0.2$ and $\beta = 4\text{dB}$ provided the best recognition rates. Tuning the latter parameter is essential to computing a mask that provides a good tradeoff between finding as many of the speech regions and not picking up too much noise.

2.2. Median filtering and blurring towards a smoother mask

In general, most of the noise variance remains in the mask after mapping the SNR to [0,1] via the sigmoid (see Fig.3.c.). The next two steps aim at refining the estimates while exploiting the spectro-temporal correlation of speech.

The first step is to apply a 3×5 two-dimensional median filter. This filter aims to erase the outliers due to the noise variability. For example, a bin with high power surrounded by a majority of lower power bins is most likely an artifact of noise variance, but will mislead the current SNR estimator into thinking it contains speech information. The median filter corrects these errors, and outputs a mask with an increased spectro-temporal coherence (see Fig.3.d.).

The second step aims to smooth the rather sharp and piecewise-constant decision regions created by the median filter. Spatial averaging is performed with a constant disk of radius 2. This smoothing serves the same purpose of noise variability cancellation as the Median filter, but acts in a complementary way, by outputting a smooth, yet well-segmented soft-decision mask (see Fig.3.e.).

Filter parameters have been optimized empirically, and will depend on the frame rate, window size and type, as well as on the number of Mel channels. In our experiments, we used a Hamming window of length 25ms that we shift by 10ms between successive frames. The number of Mel channels is 32.

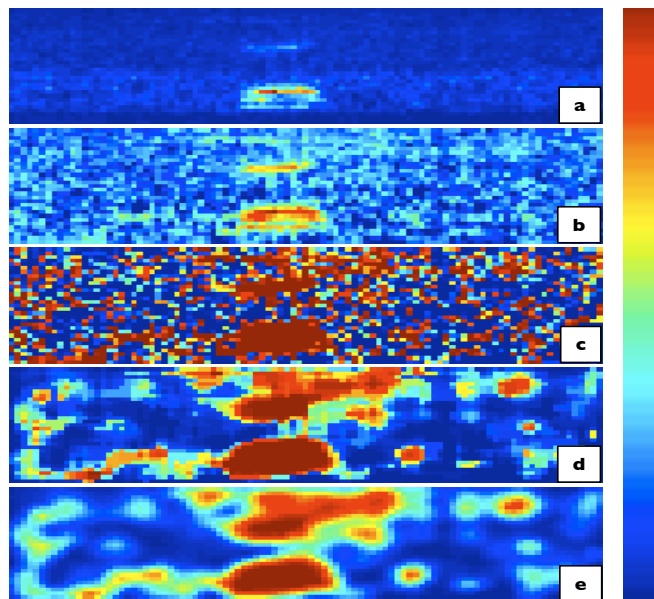


Fig.3. Output at each step of the processing of the soft mask for the digit 'SIX' corrupted by car noise at 5dB. The x-axis corresponds to time and y-axis to the Mel frequency. (a) Mel-filtered spectrum, (b) a-posteriori SNR, (c) after mapping to [0,1], (d) after median filtering and (e) the final soft mask after blurring. Note that (c, d, e) range in [0,1], while (a) and (b) both use a different scale.

2.3. Adaptive noise power estimation

To overcome the issues of varying noise conditions discussed in Section 2.1, we propose to use the efficient noise power spectral density estimator from [12], implemented as *estnoisem.m* in the Matlab Voicebox Toolkit [13]. This function outputs a frame-by-frame and channel-by-channel estimate of the noise power in the linear frequency domain. Since this estimate has a tendency to produce outliers at the beginning of the speech segment as well as sharp transitions in the estimated noise magnitude across time, we have found it useful to smooth the output by applying a median filter of length 50 frames (0.5s). The resulting noise power estimate has the desired time-smoothness of 2.1 so as to fit our mask estimation framework and automatically adapts to varying powers of noise.

After filtering the noise magnitude by a mel filter bank, we compared it with the estimate from Sec. 2.1. and noticed that, on average, it overestimates the noise magnitude by a factor of 1.5 to 2. This bias is cancelled by multiplying the latter estimate by a factor of 0.6 before computing the SNR. The following steps in computing the soft-mask are done as described in Sec. 2.1.

3. A SIMPLE APPROACH TO LOG-SPECTRAL ENHANCEMENT

In this section, we review the steps of the proposed enhancement framework. A flowchart for Section 3 can be found in Fig.2.

3.1. Soft-Mask weighting as an alternative to data imputation

The proposed algorithm is based on the idea that, if the soft mask already contains some information based on the spectro-temporal correlation of speech, then the imputation can be made quite easily. Using such a soft mask, the filled-in values of unreliable bins could originate from a weighted value of the original unreliable bin. In other words, the neighboring bins to an unreliable one help decide what proportion of its power should be retained. As an illustrative example, suppose a bin tagged as unreliable has many neighbors tagged as reliable. Then, because speech is known to be so highly time-frequency correlated, we might consider using a fraction of the noisy value of that bin, instead of setting it to zero or to some interpolated value from the neighboring reliable bins, as done in traditional imputation techniques [1] [7] [11].

With this in mind, the proposed algorithm performs enhancement of the noisy signal by simply multiplying the observed log-spectral value in each bin with its corresponding soft decision mask (Fig.4.). It should be noted that this approach has similarities with [8], where a binary mask is also used as a set of multiplicative coefficients to be applied directly on the noisy spectrum. Yet, our approach differs from [8] because we use a soft-decision mask instead of a binary mask, and that the enhancement is done in the Mel-filtered log-spectral domain as opposed to the spectral domain. To better perceive the difference with [8], note that a multiplication in the log-domain is equivalent to raising the Mel-filtered spectral amplitude to some power.

3.2. Log-Spectral flooring

The proposed approach of discarding the likely non-speech components of the Mel-filtered log-spectrum efficiently removes most of the corruptive noise. Yet, the clean log-spectra exhibits a higher dynamic range than the enhanced spectra due to the components with negative values, representative of low-energy speech and

silence. Because of our weighting with coefficients lower than 1, such values are seldom observed in the enhanced noisy spectrum.

A solution can be found in the technique of Log-Spectral Flooring [1]. We compute the liftered log-spectrum and set a flooring threshold, empirically optimized at 0dB. This step efficiently reduces the dynamic range of the log-spectrum while relying on the fact that discriminative information for ASR is more likely to be found in the peaks of the spectrum than in the valleys.

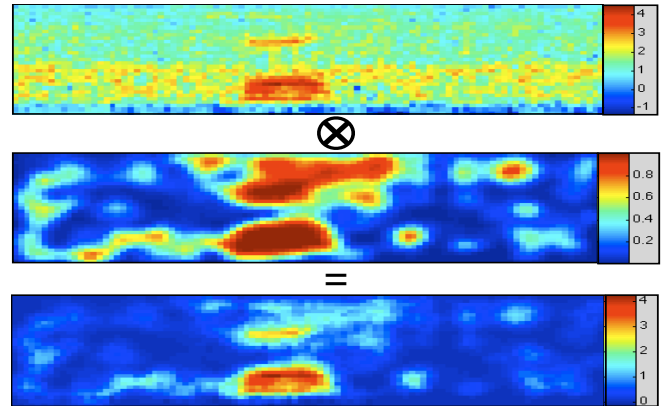


Fig. 4. Enhancement via log-spectrum weighting. The utterance is the same as in Fig.3. Noisy log Mel-spectrum (top), soft mask computed in Sec.2. (middle), enhanced log-spectrum obtained as the product of the noisy spectrum and the mask (bottom).

3.3. Gaussian smoothing

We perform smoothing directly on the log-Spectrum using a two-dimensional low-pass Gaussian filter of size 5×5 with a standard deviation of $\sigma = 0.7$ bins. This rather sharp filter helps remove the remaining noise variability on the parts of the log-spectrum that have been preserved by the mask multiplication step. This block is presented twice in the flowchart (Fig.2.): once right after multiplication by the mask, to avoid this variance to be enhanced by the liftering step and once right after the flooring, to smooth the liftered spectrum.

4. EXPERIMENTAL SETUP

Experiments have been carried out using the HTK-based back-end along with the Aurora-2 corpus of noisy digits [14]. We model 11 words with 18-state 3-mix. HMMs. Two silence models are used with, respectively, 5 and 3 states, and 3 and 6 mixtures per state.

The recognizer is trained on clean utterances only and no attempt is made to perform enhancement on the clean training set: the only pre-processing we run for MFCCs is the log-spectral flooring and the smoothing. Testing sets A and B comprise the same utterances corrupted by 8 types of background noise at various SNRs. On the testing set, we perform the full enhancement as described in Secs. 2 and 3. We evaluate the proposed method both with the oracle noise estimate from Sec. 2.1. and with the adaptive estimate from Sec. 2.3. We also run basic MFCCs, as well as MFCCs enhanced with Log-Spectral Flooring (LS-FLR) [1] alone, to emphasize the contribution of our enhancement method. We report results presented in [15] for the MVA technique, as their back-end configuration matches our setup. We also evaluated the novel PNCC features using the code from [18], which we enhanced with first and second derivatives. Finally, we show results for the ETSI advanced front-end [16], as reported in [17].

5. RESULTS AND DISCUSSIONS

In Table 1, we observe the word accuracies of the front-ends introduced above, averaged across all 8 noise types. The proposed algorithm obtains the best accuracies on this task and we notice that its two versions perform almost equally well. That is, our adaptation of the adaptive noise power estimator from [12] allows us to handle time varying noise with no significant loss of accuracy when compared to the oracle estimation proposed in II.A. The only downside lies in the increased computational load. On our machine (2.2Ghz Intel Core 2 duo MacBook Pro), the extraction of 1001 utterances with Matlab takes 40s with the oracle noise estimate of Sec. 2.1. and 95s for the adaptive estimate of Sec. 2.3. For comparison, the ETSI script in C takes about 50s on the same task while the PNCC Matlab script takes an average of 20min.

Table 1. Word-accuracies for different front-ends on Aurora-2.

SNR (dB)	20	15	10	5	0	Avg.
MFCC	97.6	93.6	78.7	45.8	11.9	65.5
LS-FLR	97.5	94.7	86.2	64.8	29.5	74.5
PNCC [18]	98.7	97.3	93.3	81.1	53.7	84.8
MVA [15]	97.9	96.1	91.6	81.0	59.2	85.1
ETSI [16]	98.1	96.7	92.8	83.2	59.8	86.1
Prop. 2.1	98.3	97.1	93.9	83.2	59.6	86.4
Prop. 2.3	98.3	97.0	93.4	82.6	58.6	86.0

Table 2. Detailed per-noise word-accuracies at 0dB SNR.

Noise	Subway	Babble	Car	Exhibition
Prop. 2.1	57.81	55.23	66.81	62.63
Prop. 2.3	53.42	49.73	68	62.45
Noise	Restaurant	Street	Airport	Train
Prop. 2.1	51.4	56.92	63.67	62.17
Prop. 2.3	48.05	63.12	62.06	61.99

6. CONCLUSION

We propose a novel framework for Mel-filtered log-spectrum enhancement, with application to noise robust ASR. First, a soft-decision mask is computed from the Mel spectrum that exploits the spectro-temporal correlation of speech by applying two simple 2D filters. A noise estimation procedure adapted from [12] is integrated into our framework, allowing it to handle varying noise conditions. Next, a simple way to enhance the signal is proposed, that uses the estimated mask as a set of multiplicative coefficients to be applied to the log-spectrum. The resulting similarity between noisy and clean MFCCs is increased, while preserving discriminative information about the speech. On the Aurora-2 task with a fixed back-end, this method is shown to perform better than state-of-the-art like MVA, PNCC or ETSI. We have been running large vocabulary ASR on Aurora-4 to assess the generalizability of our findings and found that PNCCs tend to perform better than the proposed method in that case. Future work will attempt to bridge this performance gap for a large-vocabulary setup.

ACKNOWLEDGMENT

The authors would like to thank Jonas Borgström for his valuable support and comments during the investigations that lead to this

research. This work was supported in part by the DoD RATS award via SRI.

7. REFERENCES

- [1] B. J. Borgström and A. Alwan, "Missing Feature Imputation of Log-Spectral Data For Noise Robust ASR", *Workshop on DSP in Mobile and Vehicular Systems*, 2009.
- [2] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", *IEEE TASSP*, Vol. 32, pp. 1109-1121, 1984.
- [3] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Log-Spectral Amplitude Estimator", *IEEE TASSP*, Vol. 33, No. 2, pp. 443-445, 1985.
- [4] B. J. Borgström and A. Alwan, "Improved Speech Presence Probabilities Using HMM-Based Inference, With Applications to Speech Enhancement and ASR", *IEEE Journal of Selected Topics in Signal Processing*, vol.4, no.5, pp.808-815, Oct. 2010
- [5] B. J. Borgström and A. Alwan, "A Statistical Approach to Mel-Domain Mask Estimation for Missing-Feature ASR", *IEEE Signal Processing Letters*, Vol. 17, No. 11, pp. 941-944.
- [6] J. Barker, M. Cooke and P. Green "Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise", *EuroSpeech 2001*, pp. 213-216.
- [7] B. Raj, M. L. Seltzer, and Richard M. Stern, "Reconstruction of missing features for robust speech recognition", *Speech Communication*, vol. 43, pp. 275-296, 2004
- [8] W. Hartmann and E. Fosler-Lussier, "Investigations into the Incorporation of the Ideal Binary Mask in ASR", *ICASSP 2011*, pp. 4804-4807.
- [9] J. F. Gemmeke, H. Van Hamme, B. Cranen, and L. Boves, "Compressive sensing for missing data imputation in noise robust speech recognition", *IEEE J. Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 272-287, 2010
- [10] J. F. Gemmeke, T. Virtanen and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition", *IEEE TASLP*, vol.19, no.7, pp.2067-2080, Sept. 2011
- [11] B. J. Borgström and A. Alwan "Utilizing Compressibility in Reconstructing Spectrographic Data, with Applications to Noise Robust ASR", *Signal Processing Letters*, vol. 16, Issue 5, pp. 398-401, 2009
- [12] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics", *IEEE TASP*, vol. 9, pp.504 - 512, 2001.
- [13] M. Brookes, "Voicebox, Speech Processing Toolbox for MATLAB", Department of EE, Imperial College, London, www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
- [14] D. Pearce and H. G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", *Automatic Speech Recognition: Challenges For the New Millennium*, ASR2000, 2000, pp. 181-188.
- [15] C. Chen, J. Bilmes and K. Kirchoff, "Low-Resource Noise-Robust Feature Post-Processing on Aurora 2.0", *ICLSP 2002*, pp. 2445-2448.
- [16] Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Adv. Frontend Feature Extraction Algorithm; Compression Algorithms, *ETSI ES 202 050*, 2007.
- [17] V. Mitra, H. Nam, C.Y. Espy-Wilson, E. Saltzman and L. Goldstein, "Articulatory Information for Noise Robust Speech Recognition", *IEEE TASLP*, vol.19, no.7, pp.1913-24, Sept. 2011.
- [18] C. Kim and R.M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring", *ICASSP 2010*, pp. 4574-4577.