SPEECH ENHANCEMENT USING A PITCH PREDICTIVE MODEL

Luis Buera

Communication Technologies Group (GTC)
University of Zaragoza, Spain
lbuera@unizar.es

Jasha Droppo and Alex Acero

Speech Research Group
Microsoft Research, Redmond, WA, USA
{jdroppo,alexac}@microsoft.com

ABSTRACT

In this paper we present two new methods for speech enhancement based on the previously publised fine pitch model (FPM) for voiced speech. The first method (FPM-NE) uses the FPM to produce a non-stationary noise estimate that can be used in any standard speech enhancement system. In this method, the FPM is used indirectly to perform speech enhancement. The second method we describe (FPM-SE) uses the FPM directly to perform speech enhancement. We present a study of the behavior of the two models on the standard Aurora 2 task, and demonstrate improvements of over 45% average word error rate reduction over the multi-style baseline.

Index Terms— Speech enhancement, Speech analysis, Speech recognition, Robustness

1. INTRODUCTION

Modeling speech signals is a fundamental problem in many speech processing applications. The most common models assume that the speech can be represented by an excitation signal filtered by a linear model, which represents the vocal tract. Note that pitch, which comes from the repeated closing of the glottal folds within the larynx, is an important piece in this model due to most of the energy of voiced speech segments is concentrated in pitch fundamental frequency and its harmonics.

The pitch information has been used in many speech applications, such as coding [1], speaker identification [2], speech enhancement [3], and robust speech recognition [4], obtaining important benefits.

In this paper, we propose to use the pitch period to build both noise (FPM-NE) and speech (FPM-SE) long term models for human speech enhancement and robust speech recognition. The noise model is based on a time varying comb filter, while the speech model consists on two additive terms: the pitch period shifted version of the speech and a term which represents the non-voiced segments. In both cases, the pitch period estimation is fundamental and it is computed with a fine pitch tracker [5].

To study the behavior of the two predictive models, some experiments with Aurora 2 corpus [6] were carried out, obtaining important noise reduction and interesting ASR improvements in both cases: 49.67% of average improvements in FPM-NE, where the noise model is used jointly with VTS enhancement [7], and 50.31% in FPM-SE, where the speech prediction model is applied directly.

This paper is structured as follows. In Section 2, a brief overview of the fine pitch tracker is included. In Sections 3 and Section 4, the noise and speech models based on pitch period are presented respectively. In Section 5, the behavior of the two models are studied. Finally, the conclusions and lines of future are included in Section 6.

2. FINE PITCH TRACKER

The fine pitch tracker, first introduced in [5], is different from most pitch estimation algorithms in that it operates time-synchronously with the incoming speech signal, $y(n)$. It does this by estimating a pitch track $\tau(n)$ that minimizes the objective function $\mathcal{F}$.

$$\mathcal{F} = \sum_n (y(n) - y(n - \tau(n))^2) + \beta \sum_n (\tau(n) - \tau(n-1))^2. \tag{1}$$

Note that $\mathcal{F}$ is composed of two terms. The first term measures the energy of the residual error between $y(n)$ and a long term model based on pitch period. The second term introduces a penalty when the pitch period changes from one sample to the next. The parameter β controls the relative importance between these two terms.

Because $\mathcal{F}$ is a first-order Markov in $\tau(n)$, the optimum instantaneous pitch sequence can be found using standard dynamic programming search techniques.

3. FPM NOISE ESTIMATION

It is common practice to model environmental noise as additive distortion in the time domain, as in Eq. (2). Here, the hypothetical clean speech $x(n)$ has been corrupted by the additive noise $w(n)$ to produce the noisy observation $y(n)$.

$$y(n) = x(n) + w(n), \tag{2}$$

Voiced speech energy is concentrated at the pitch frequency and in its harmonics. Outside of these frequencies, the noise can be estimated with a time varying comb filter as in Eq. (3). Here, $\hat{w}(n)$ is the noise estimate, and $\hat{\tau}(n)$ is a fine pitch track valid for all values of n.

$$\hat{w}(n) = y(n) - y(n - \hat{\tau}(n)), \tag{3}$$

Although this time-varying notch filter is very effective at removing the voiced speech, there are two remaining problems that must be solved.

The first problem is that although the voiced speech has been removed at the harmonics, the noise energy at these frequencies has also been eliminated. Only frequencies in-between these harmonics represent good estimates of the noice.

The second problem is that in addition to noise, the estimate $\hat{w}(n)$ will contain unvoiced speech energy, such as fricitives and plosives.

To address both of these problems, a time-varying spectral representation is created from $\hat{w}(n)$ using a short-time Fourier transform. This representation is then smoothed in both time and frequency. The resulting sequence of nonstationary noise spectral estimates can then be used with any basic enhancement method. In this work, we have chosen the technique Vector Taylor Series, VTS, for feature vector enhancement [7].

4. FPM SPEECH ESTIMATION

In order to obtain the speech estimation with the fine pitch model, two assumptions and a training process with synthesizing data are needed.

First, we assume that the acoustic environment is modeled by additive noise (as in Eq. 2), where $w(n)$ is a white Gaussian process with 0 mean and variance $\sigma_w^2(n)$.

Second, we assume that $x(n)$ can be predicted by

$$x(n) = a(n)x(n - \tau(n)) + v(n). \quad (4)$$

Here, the modulation factor $a(n)$ measures the relative importance of the long term prediction model. Note that $a(n)$ should be near 1 in perfectly periodic regions and near 0 in unvoiced or silence regions. Furthermore, its variation should be smooth.

The term $v(n)$ represents the prediction error of the speech long term model, and takes the form of a zero-mean, $\sigma_v^2(n)$ variance, white Gaussian process. This term should be very small in silence and purely voiced regions, and larger in unvoiced segments when the long term speech model is inaccurate.

4.1. Estimation of the clean signal

Combining (2) and a Gaussian prior for $w(n)$, the probability of noisy signal, $y(n)$, given the clean one, $x(n)$, and the variance of the additive noise, $\sigma_w^2(n)$, $p(y(n)|x(n), \sigma_w^2(n))$, can be computed as

$$p(y(n)|x(n), \sigma_w^2(n)) = \mathcal{N}(y(n); x(n), \sigma_w^2(n)). \quad (5)$$

As well, $p(x(n)|a(n), x(n - \tau(n)), \sigma_v^2(n))$, which is the probability of the clean speech signal, $x(n)$, given the long term speech model and $\sigma_v^2(n)$ can be obtained as

$$p(x(n)|a(n), x(n - \tau(n)), \sigma_v^2(n)) = \mathcal{N}(x(n); a(n)x(n - \tau(n)), \sigma_v^2(n)). \quad (6)$$

Finally, the estimation of the clean speech signal, $\hat{x}(n)$, is obtained maximizing the joint probability of $x(n)$ and $y(n)$, $p(y(n), x(n)|x(0), ..., x(n-1), \sigma_w^2(n), \sigma_v^2(n))$, with respect to $x(n)$. Observe that this expression can can be computed combining (5) and (6)

$$\hat{x}(n) = \gamma(n)y(n) + (1 - \gamma(n))a(n)\hat{x}(n - \tau(n)). \quad (7)$$

$$\gamma(n) = \frac{\sigma_v^2(n)}{\sigma_v^2(n) + \sigma_w^2(n)}. \quad (8)$$

Thus, $\hat{x}(n)$ is composed by two terms weighted by $\gamma(n)$ and $(1 - \gamma(n))$. For time samples with high SNR, the energy of the prediction error of the speech long term model will be bigger than the energy of the additive noise ($\sigma_v^2(n) >> \sigma_w^2(n)$). As a result, $\gamma(n)$ will be approach the value 1, and Eq. (7) reduces to $\hat{x}(n) \approx y(n)$. On the other hand, if $\sigma_v^2(n) << \sigma_w^2(n)$, $\gamma(n)$ will go to zero, and the most important term will be the second one, which represents the estimated speech from the long term predictive model.

Note that, because $\gamma(n)$ depends on the SNR ($\sigma_w^2(n)$) and the nature of the speech signal ($\sigma_v^2(n)$), it should vary smoothly over time.

Assuming that $\tau(n)$ is obtained as Section 2, we still need to estimate $\gamma(n)$ and $a(n)$, which is described in the following sections.

4.2. Estimation of $\gamma(n)$

To estimate $\gamma(n)$, we build a model that learns the joint distribution of γ sequences and given noisy speech observations. This model is trained with synthetic stereo data, and then applied to produce a MMSE estimate of γ given the noisy speech observations.

4.2.1. Synthesizing training data

To train the joint distribution of γ sequences and noisy speech observations in acoustic environment e, we need training data that contains matched pairs of noisy speech and optimal γ sequences $\{y_e(n), \gamma(n)\}$. We synthesize this training data from an existing set of data that contains match pairs of noisy and clean utterances $\{y_e(n), x_e(n)\}$.

Given a matched pair of clean and noisy stereo training data, optimal sequences $\gamma(n)$ and $a(n)$ are jointly estimated by minimizing the MSE (Eq. 9) between the clean training signal, $x_e(n)$, and the corresponding estimation, $\hat{x}_e(n)$ obtained from (Eq. 7).

$$\gamma_{e,opt}(n), a_{e,opt}(n) = \underset{\gamma(n),a(n)}{arg\,min} \sum_n (x_e(n) - \hat{x}_e(n))^2. \quad (9)$$

In general, several options can be considered to minimize an objective scalar function of several variables, but in this work a gradient minimization algorithm with some smoothness constraints over $a(n)$ and $\gamma(n)$ is used. The smoothness constraints consist on assuming that $a(n)$ and $\gamma(n)$can be built as an addition of several low frequency sin functions. Note that gradient minimization algorithms are in theory suboptimal, nonetheless satisfactory solutions are obtained.

4.2.2. Training process

Given stereo synthesized training data $\{x_e(n), y_e(n), \gamma_{e,opt}(n), a_{e,opt}(n)\}$, we assume that noisy MFCC feature vectors, $\mathbf{y}_e^{mfcc}$, can be modeled following a GMM. Observe that this is a reasonable way to split the noisy space because $\gamma(n)$ depends on SNR and the characteristics of the speech.

$$p(\mathbf{y}_e^{mfcc}) = \sum_{s_y^e} p(\mathbf{y}_e^{mfcc}|s_y^e)p(s_y^e), \quad (10)$$

$$p(\mathbf{y}_e^{mfcc}|s_y^e) = \mathcal{N}(\mathbf{y}_e^{mfcc}; \mu_{s_y^e}, \mathbf{\Sigma}_{s_y^e}), \quad (11)$$

where $\mu_{s_y^e}$, $\mathbf{\Sigma}_{s_y^e}$ and $p(s_y^e)$ are the mean vector, the covariance matrix and the a priori probability of the noisy model Gaussian s_y^e. Thus, the value of γ associated to s_y^e, $\gamma_{s_y^e}$, is

$$\gamma_{s_y^e} = \frac{\sum_n p(s_y^e|y_e(n), e)\gamma_{e,opt}(n)}{\sum_n p(s_y^e|y_e(n), e)}, \quad (12)$$

where $p(s_y^e|y_e(n),e)$ is the a posteriori probability of the noisy model Gaussian s_y^e, given the noisy sample $y_e(n)$ and the e basic environment. This probability can be computed with the corresponding MFCC feature vector associated to $y_e(n)$, $\mathbf{y}_{e,n}^{mfcc}$, as

$$p(s_y^e|y_e(n),e) = \frac{p(\mathbf{y}_{e,n}^{mfcc}|s_y^e)p(s_y^e)}{\sum_{s_y^e} p(\mathbf{y}_{e,n}^{mfcc}|s_y^e)p(s_y^e)}. \quad (13)$$

4.2.3. MMSE estimator

Given the testing noisy signal, $y(n)$, the estimation of $\gamma(n)$ is obtained with MMSE criterion as

$$\hat{\gamma}(n) = \sum_e \sum_{s_y^e} p(e|y(n))p(s_y^e|y(n),e)\gamma_{s_y^e}, \quad (14)$$

where $p(s_y^e|y(n),e)$ is the a posteriori probability of the noisy model Gaussian s_y^e, given the noisy basic environment and the noisy signal, $y(n)$. This expression can be obtained in a similar way as (13) using the corresponding MFCC feature vector $\mathbf{y}_n^{mfcc}$ associated to $y(n)$. On the other hand, $p(e|y(n))$ is the a posteriori probability of the basic environment e given $y(n)$, which can be obtained using (10) and (11) as

$$p(e|y(n)) = \varphi p(e|y(n-1)) + (1-\varphi)\frac{p(\mathbf{y}_n^{mfcc})}{\sum_e p(\mathbf{y}_n^{mfcc})}, \quad (15)$$

where φ is the memory term (0.98 in this work), and $p(e|y(0))$ is considered equiprobable for all the basic environments.

4.3. Estimation of $a(n)$

Although several objective functions have been considered in order to estimate $a(n)$, in this work we have chosen the following function:

$$\mathcal{G} = \sum_n (y(n) - a(n)y(n-\hat{\tau}(n)))^2 + \phi(|a(n) - arg\,\underset{a_i}{max}(p(a_i|\hat{\gamma}(n)))|). \quad (16)$$

It can be observed that $\mathcal{G}$ is composed of two terms. The first one measures the energy of the residual error between $y(n)$ and the proposed long term speech model. The second term includes a penalty between $a(n)$ and the discrete a priori most probable value of $a(n)$, a_i, given $\hat{\gamma}(n)$, which has been previously estimated as (14). So, the probability of a_i, given $\hat{\gamma}(n)$, $p(a_i|\hat{\gamma}(n))$, can be estimated as

$$p(a_i|\hat{\gamma}(n)) = \sum_e \sum_{s_y^e} p(e|y(n))p(s_y^e|y(n),e)p(a_i|\hat{\gamma}_i, s_y^e), \quad (17)$$

where $\hat{\gamma}_i$ is the discretized value of $\hat{\gamma}(n)$ and $p(a_i|\hat{\gamma}_i, s_y^e)$ is the probability of a_i, given $\hat{\gamma}_i$ and the noisy Gaussian s_y^e. Note that $p(a_i|\gamma_i, s_y^e)$ can be estimated in the training process with the synthesizing training data. The penalty function, $\phi(\epsilon)$, can be chosen in different ways. In this work we have assumed an infinity penalty when $\epsilon > 0.2$, and zero in other case. Finally, a gradient minimization algorithm with some smoothness constraints over $a(n)$ is used to minimize $\mathcal{G}$. The smoothness constraints consist on assuming that $a(n)$ can be built as an addition of several low frequency sin functions.

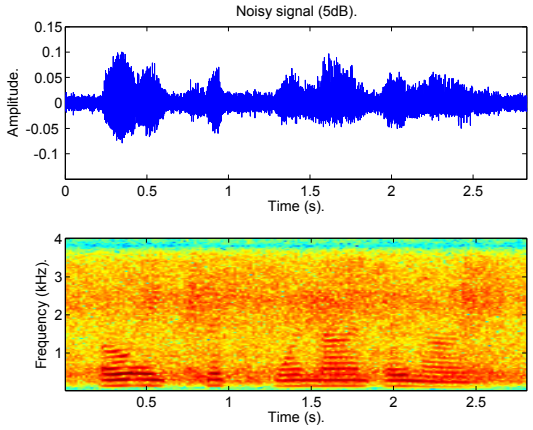

Fig. 1. Testing noisy utterance in time and frequency domains (5dB SNR, subway noise, set A).

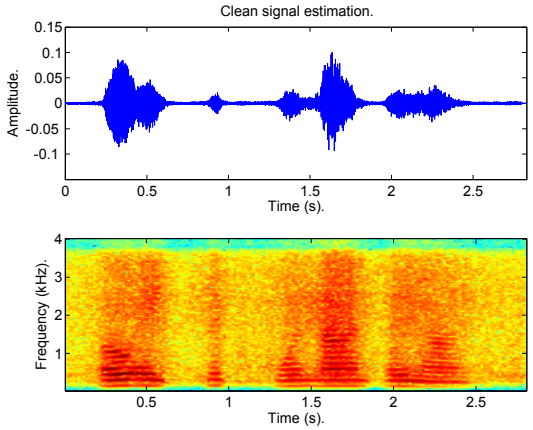

Fig. 2. Enhanced testing utterance in time and frequency domains when the proposed FPM noise estimation model is used with VTS feature enhancement.

5. RESULTS

To study the performance of the two proposed models, a set of experiments were carried out using Aurora 2 database [6].

The Aurora 2 task is isolated and continuous digits. As feature set, the standard ETSI front-end [8] features plus energy and the corresponding delta and delta delta coefficients are used. Whole-utterance cepstral mean normalization is applied to testing and training data. The acoustic models are composed of 16 state HMM for each digit, a 3 state begin-end silence HMM and a 1 state inter-word silence HMM. In all cases, each pdf state is composed by a mixture of three Gaussians.

In training the parameters for the FPM Speech estimation model, identical utterances from the clean training set and the multicondition training set were used. In effect, the speech model is tuned on the noise types from set A, keeping the noise types from sets B and C as unseen conditions. Also, the results for "clean condition" training actually use the multicondition data.

In the FPM Speech estimation model, noisy MFCC feature vectors are modeled as a GMM on static cepstral features with 32 components. In the FPM Noise estimation model, the clean speech GMM consisted of a 32 component GMM on static cepstral features.

Imp.(%)	-5dB	0dB	5dB	10dB	15dB	20dB	Clean	Ave.
Multi NE	15.65	35.73	36.48	36.51	40.37	42.45	40.88	37.56
Clean NE	17.07	49.12	68.60	72.74	63.49	41.44	-26.41	61.78
Multi SE	19.18	40.53	47.61	54.27	61.31	68.46	73.37	45.88
Clean SE	8.52	33.43	62.92	75.49	75.69	73.18	63.25	54.74

Table 1. Average improvements obtained with Aurora 2 database for the proposed predictive models (Noise Model with VTS, "NE", and Speech Model, "SE") and different training conditions: multi condition, "Multi", and clean, "Clean".

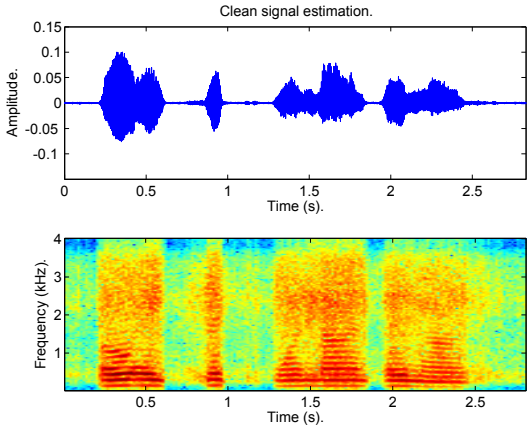

Fig. 3. Enhanced testing utterance in time and frequency domains when the proposed FPM speech enhancement model is used.

In Fig. 1, one testing noisy utterance (5dB SNR, subway noise, set A) is plotted in time and frequency domains. The corresponding enhanced signals with the FPM noise and speech enhancement models are included in Fig. 2 and 3, respectively. An important noise reduction without distortion can be appreciated in both cases, although it is more significant in the second one (note an amplitude reduction of the speech segments in Fig. 2 due to a poor estimation of the noise). Since important noise reduction has been observed for different SNR, we can assert that the proposed speech model provides a satisfactory human being enhancement performance for a significant range of SNR.

The average ASR improvements obtained with Aurora 2 database are presented in Table 1, where "NE" indicates that the noise predictive model is used jointly VTS for feature vector enhancement, while "SE" represents that the enhanced utterances are computed with the speech predictive model. The different training conditions (clean and multicondition) are also included, "Clean" and "Multi", respectively. It can be observed that the technique based on the noise predictive model produces better results in low SNR environments with clean training conditions, while the performance obtained with the speech predictive model is quite better in medium and high SNR, which are the most important for real applications (more than 70% in average with 10dB, 15dB, 20dB and clean). However, the results with multicondition training show a better behavior in all situations when the speech predictive model is used.

6. CONCLUSIONS AND FUTURE WORK

In this paper, novel noise and speech predictive models have been presented. The FPM-NE model is based on a time varying comb filter and uses a non-stationary noise estimator jointly with VTS speech enhancement. The FPM-SE model uses the fine pitch model, together with conditional priors for the enhancement parameters $\gamma(n)$ and $a(n)$, to directly enhance the speech signal.

Interesting speech recognition results have been obtained in both cases against the Aurora 2 database. The best results are from the FPM-SE model, which achieves a 50.91% WER reduction on average, and more than 70% WER reduction in SNR conditions above 10dB. These high SNR conditions are the most interesting for building real applications.

The FPM-SE algorithm presented in this paper could be improved in two significant ways. First, the presented method of computing $\gamma(n)$ and $a(n)$ from the noisy signal is rather ad-hoc and could be replaced with something more principled using a time-domain speech model. Second, where we assume the nature of $\sigma_v^2(n)$ and $\sigma_w^2(n)$ can be captured in a single variable $\gamma(n)$, there are potential benefits to separately modelling these two variances.

7. REFERENCES

[1] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*, Prentice Hall PTR, 2001.

[2] H. Ezzaidi, J. Rouat, and Douglas O'Shaughnessy, "Towards combining pitch and mfcc for speaker identification systems," in *Proc. Eurospeech*, 2001, pp. 2825–2828.

[3] A.-T Yu and H.-C. Wang, "New speech harmonic structure measure and its application to post speech enhancement," in *Proc. ICASSP*. IEEE, 2004, vol. I, pp. 729–732.

[4] M. Seltzer, J. Droppo, and A. Acero, "A harmonic-model based front end for robust speech recognition," in *Proc. Eurospeech*, 2003, pp. 1277–1280.

[5] J. Droppo and A. Acero, "A fine pitch model for speech," in *Proc. Interspeech*, 2007.

[6] H. G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *Proc. in ISCA ITRW ASR2000*, Paris, France, September 2000.

[7] D. Y. Kim, C. K. Un, and N. S. Kim, "Speech recognition in noisy environments using first-order vector taylor series," *IEEE Transactions on Signal Processing*, vol. 5, no. 3, pp. 57–59, March 1998.

[8] ETSI, "Speech processing transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms," Tech. Rep., ETSI ES 201 108 version 1.1.2, April 2000.