ISCA Archive
http://www.isca-speech.org/archive

EVALUATION OF A PITCH ESTIMATION ALGORITHM FOR SPEECH EMOTION RECOGNITION

N. Vanello[1], N. Martini[2], M. Milanesi[3], H. Keiser[1], M. Calisti[4], L. Bocchi[4], C. Manfredi[4], L. Landini[1]
[1]Department of Information Engineering, University of Pisa, Pisa, Italy [2]Interdepartmental Research Center "E. Piaggio", University of Pisa, Pisa, Italy [3]MRI Laboratory, "G. Monasterio" Foundation, Pisa, Italy [4]Department of Electronics and Telecommunications, University of Florence, Florence, Italy

Abstract: **The analysis of parameters extracted from speech data may contribute, together with other approaches, to the analysis and classification of a subject emotional status. Pitch value and variability have been shown to carry useful information to reach this goal. However the non stationarity of running speech and the short duration of utterances represent a difficulty for the estimation of these parameters. In this work a method based on a variation of the Sawtooth Waveform Pitch Estimator (SWIPE') to estimate pitch and jitter in vowel sound, is evaluated. The performances of the approach are assessed on simulated datasets with varying signal to noise ratios and jitter values. Issues related to data length are introduced and discussed through simulations. A comparison of the approach performances with the Simplified Inverse Filtering Technique (SIFT) is presented. Preliminary results on vowels extracted from a database of emotional utterances are introduced.**

Keywords : **Pitch, jitter, swipe', emotion, vowels**

I. INTRODUCTION

The development of automatic methods to estimate subjects' psychological status has drawn the attention of the research community. The achievement of such information has several positive outcomes on fields such as psychology, for development of tools for patients monitoring or for improving occupational safety. To reach this goal multiparametric approaches have been proposed as those based on the acquisition of vital signs related to the activity of the autonomous and the central nervous systems and on the analysis of speech. As regards the latter approach several features have been proposed as those based on speaking rate, spectral characteristics and prosody [1][2]. Pitch related variables have been proposed as F0 level, range, contour and jitter. In particular F0 mean values and variability was found to be larger for angry and happy speech rather than neutral or sad speech [1].

The estimation of pitch represents a challenging task in running speech given the short duration of sounds and due to the noise [3]. Moreover the non stationarity of speech signals requires the use of short analysis windows thus allowing to estimate the changes of pitch across time.

In this work a method based on a variation of the Sawtooth Waveform Inspired Pitch Estimator, namely SWIPE' algorithm [5], is introduced for pitch and jitter estimation in vowels sounds. The approach is tested by using synthesized vowels and results are compared with those obtained by the Simplified Inverse Filtering Technique (SIFT) [6]. Application for classification of vowels as extracted by emotional utterances is introduced.

II. METHODS

Synthetic data were obtained by an autoregressive moving average exogenous (ARMAX) model. The parameters of the model were estimated from an healthy male /a/ vowel, with model orders for the AR, MA and X part equal to 16, 4 and 2 respectively. The model input for synthesis purposes was obtained with an impulse train sequence, whose distance between two successive pulses was modulated to produce the desired jitter. The amount of the imposed jitter was changed across different simulations ranging from a minimum of 0 to a maximum of 2 percent. The signal to noise ratio (SNR) of the simulated vowels was modified by inserting additive Gaussian noise at the model output.

Real dataset consisted of vowels extracted from a German database of emotional utterances [7]. Ten different sentences are repeated by different actors and labelled according to perceived emotional content, respectively as neutral, anger, fear, joy, sadness, disgust and boredom. Vowels were extracted from the sentences according to dataset labels and segmentation provided with the datasets. Dataset labeling and segmentation is based on auditive judgement supported by visual analysis of oscillogram and spectrogram, as described in [7].

SWIPE' algorithm measures pitch by estimating average peak to valley distance at harmonic locations. This goal is achieved by comparing the spectrum of the signal with that of cosine based kernel functions, thus weighting the pitch candidate and its harmonics according to a $1/\sqrt{f}$ law. This choice matches the decay trend of harmonics relative to vowels sounds. To avoid subharmoics of pitch being estimated as the real pitch, non prime harmonics, except the first one, are removed

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 6th international workshop: December 14-16, 2009*, ISBN 978-88-6453-094-9 (print), 978-88-6453-096-3 (online), © 2009 Firenze University Press.

from the kernel. This algorithm uses a window size related to the pitch to be estimated: in particular a Hann window size of length $T=4/f_0$ is chosen. We applied SWIPE' algorithm to a sliding window of length T seconds. A pitch value is estimated at every step. The algorithm results are evaluated for two different values of the window time shifts, namely T seconds, obtaining non overlapping windows, and T/4 seconds. This approach requires a two-step process, the first being a preliminary estimation of the pitch value. The time window length is then determined as $T=4/(\alpha f_0)$ where $\alpha<1$ is used to guarantee a sufficient time window length in the case of pitch underestimation after the first step. In this work $\alpha=0.9$ has been used. The pitch value is estimated as the mean value across windows. Jitter was estimated according to the following formula

$$Jitter = \frac{1}{N-1}\sum_{i=1}^{N-1}|F_{i+1}-F_i| \Big/ \frac{1}{N}\sum_{i=1}^{N}F_i \qquad (1)$$

where F_i is the estimated pitch at the i-th window. As a comparison, the same procedure was applied using the well known SIFT algorithm [7]. SIFT algorithm is based on inverse filtering of the speech data, where the filter is obtained by inverting a low order linear predictor that models the vocal tract. The pitch is then estimated by computing the autocorrelation function of the residuals that are related to the exciting source of the vocal tract.

III. Results

A. Simulated Data

In Fig. 1 (upper window) the percentage error of the estimated pitch, with respect to actual pitch, is shown as a function of SNR by using the SWIPE' based approach.

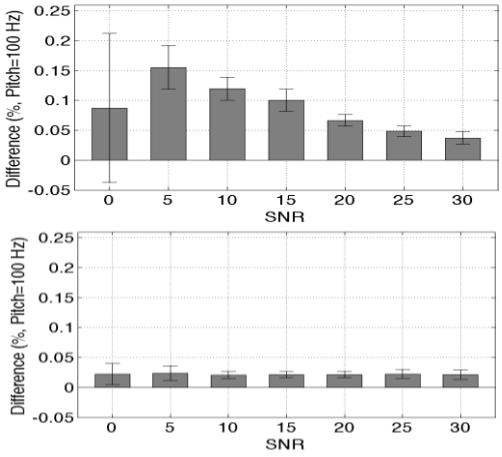

Figure. 1 Percentage error of estimated pitch at 100 Hz for different SNR values using SWIPE' and SIFT based approach (upper and lower window respectively).

Jitter estimation results for SNR=18 dB are reported in Fig. 2 to Fig. 4. The mean and the standard deviation of the estimated jitter are shown with respect to imposed jitter. For each value of the imposed jitter 20 different data segment were analyzed, each 300 ms long. In Fig. 2 the results obtained by using non overlapping windows are shown, for SWIPE' and SIFT based approach. The two algorithms yield similar results with small differences: in particular SWIPE' based approach is more accurate than SIFT based at lower jitter values and less accurate for higher jitter values.

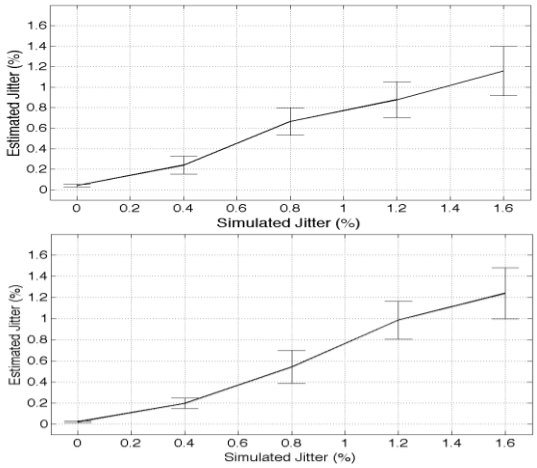

Figure 2 Estimated against imposed jitter, using SWIPE' (upper window) and SIFT (lower window) based approach and non overlapping windows. Total signal length: 300 ms.

In Fig. 3 the results obtained by using overlapping windows are shown. In this case the estimated jitter standard deviation is smaller than that obtained by employing non overlapping windows. The mean value of the estimated jitter found is always monotonically increasing with the imposed jitter. The results obtained with overlapping windows are less accurate than those obtained with non overlapping windows, resulting in an underestimation of the jitter.

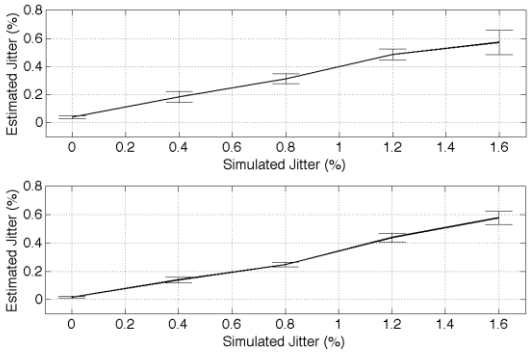

Figure 3 Estimated against imposed jitter, using SWIPE' (upper window) and SIFT (lower window) based approach and T/4 overlapping windows. Total signal length: 300 ms.

In Fig. 4 the results obtained by applying the proposed approach to a simulated vowel, 80 ms long, are shown. For lower jitter values the results obtained with the SWIPE' based approach are slightly more accurate. For higher jitter values the standard deviation of the estimated jitter is high. In particular SWIPE' based approach results in a very high standard deviation with respect mean jitter value.

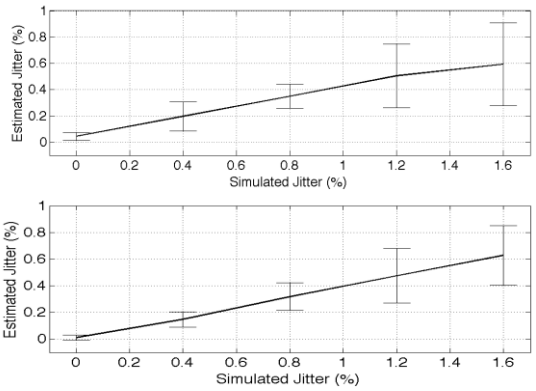

Figure 4 Estimated against imposed jitter, using SWIPE' (upper window) and SIFT (lower window) based approach and T/4 overlapping windows. Total signal length: 80 ms.

The standard deviation of the shown results is given by two sources, the estimation error of the algorithm and the trial by trial changes in the actual simulated jitter. In fact the jitter is simulated by imposing the standard deviation of the intervals between two successive pulses, given as input of the ARMAX model. By analyzing short duration windows, the actual jitter may be significantly different from the average jitter. The evaluation of the jitter from the pulse sequence used for the simulations, in fact resulted in standard deviation values equal to 0, 0.04, 0.09, 0.15 and 0.22 % for average jitter equal to 0, 0.4, 0.8, 1.2 and 1.6 % respectively. These values were estimated from the ARMAX input pulses using an 80 ms data segment and overlapping windows. In Fig. 5 the relationship between the expected jitter and the jitter estimated by the proposed approach, using SWIPE' (left) and SIFT (right) algorithm respectively, are shown.

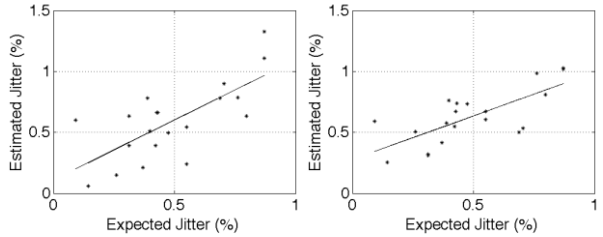

Figure 5 Regression model between estimated and expected jitter (imposed jitter=1.6%) using SWIPE' (left) and SIFT (right) based approaches (T/4 overlap).

These results pertain 20 data segment, 80 ms long, obtained imposing jitter 1.6% (see Fig. 4). The SWIPE' based approach in this simulation outperforms SIFT based one, in estimating actual jitter. A regression model is estimated describing the relationship between estimated and expected jitter. The r^2 statistic, the intercept and the slope for the linear regression models shown in the right (SWIPE' based) and the left (SIFT based) of Fig. 5 are respectively r^2=0.51, a=0.11, b=0.98 and r^2=0.53, a=0.28, b=0.71. The significance of the regression model was found to be reduced for lower imposed average jitters, resulting in lower variance estimates.

B. Real utterances

Results obtained from real dataset are shown in tables 1 and 2, for SWIPE' and SIFT based methods respectively. The results were obtained by applying the two approaches to the same vowels. Each vowel was extracted from a different subject. The results shown in Tables 1 and 2 are obtained using overlapping windows. Given the short time course of real utterances the proposed approach shows some limitations given the window length needed for pitch estimation. These preliminary results on real datasets highlight an increase of pitch values of vowels scored as anxiety and anger with respect to boredom and neutral. As regards jitter values an increase was observed predominantly in vowels scored as anger.

Table 1. Pitch (p in Hz) and jitter values (%) as estimated from real utterances using SWIPE' and overlapping windows.

Vowel	Neutral	Anxiety	Anger	Boredom
/e/	p=136 jitt=1	p=170 jitt=0.45	p=220 jitt=1.89	p=132 jitt=0.5
/i/	p=109 jitt=0.1	p=144 jitt=0.63	p=250 jitt=2.15	p=118 jitt=2.9
/u/	p=140 jitt=0.8	p=250 jitt=1.6	p=248 jitt=1.3	p=113 jitt=0.5
/a/	p=115 jitt=0.6	p=125 jitt=0.9	p=156 jitt=0.27	p=140 jitt=0.16

Table 2. Pitch (p in Hz) and jitter values (%) as estimated from real utterances using SIFT and overlapping windows.

Vowel	Neutral	Anxiety	Anger	Boredom
/e/	p=138 jitt=0.87	p=170.2 jitt=0.6	p=217 jitt=2.27	p=134 jitt=0.59
/i/	p=110 jitt=0.49	p=144.8 jitt=1.03	p=247 jitt=2.24	p=117.9 jitt=1.75
/u/	p=140 jitt=0.8	p=250 jitt=1.6	p=248 jitt=1.3	p=113 jitt=0.5
/a/	p=115.9 jitt=0.67	p=123 jitt=1.27	p=155 jitt=0.51	p=141 jitt=0.22

More severe problems are found using the approach with non overlapping windows, since the jitter estimation in case of low pitch short utterances may not be possible. The results obtained with SWIPE' and SIFT based approach are similar as regards pitch estimation, while jitter could not be estimated in most of the cases (data not shown).

IV. Discussion

Pitch estimation with the proposed method was achieved with an error smaller than 0.2% in the worst case (SNR=0) and improves considerably at higher SNRs. The comparison with SIFT based approach highlights that that the latter approach offers better results. However comparison in a wider frequency range is not explored in this work. We have to stress that a comparison of the original SWIPE' algorithm with other approaches can be found in [5], while the results in this work are related to a different approach, described in the Methods section, that may take advantage of SWIPE' as well as other pitch estimation algorithms. Given this observation this work does not aim at evaluating the SWIPE' algorithm per se but to evaluate a SWIPE' based approach for pitch and jitter estimation in short time vowels.

Jitter estimation resulted in an estimated mean value monotonically increasing with imposed jitter. The jitter values are underestimated and the average slope of the obtained results reduces in the case of overlapping windows. These results are in good agreement with those expected, given the fact that pitch is estimated by using 4/f0 seconds long windows. This choice implies an average of the pitch changes across 4 glottal cycle repetitions resulting in a systematic underestimation of the real jitter value. In the case of overlapping windows smaller changes in pitch estimation are to be expected resulting in smaller value calculated as in (1). Since our final aim is to look at possible changes of these values with respect to the expressed emotions, this issue may not represent a limitation. Moreover by using overlapping windows it is possible to give an estimate of jitter value in shorter utterances. The need for the SWIPE' algorithm to have a 4 period long time window in order to have an optimal estimate, may impose severe limits to jitter estimation for short utterances characterized by a low pitch value. In fact for 100Hz mean pitch value, a data window 40 ms long is needed. In this conditions, if the overlapping windows approach is used, a 60 ms data length would allow to estimate 3 pitch values.

By analyzing the relationship between expected and estimated jitter for higher values of the imposed jitter, a linear regression model was found to be significant or close to significance. This result shows that a large part of the jitter variance in Fig. 5 can be explained as trial by trial jitter variance in short simulated dataset. Furthermore, as it could be drawn from results in Fig. 5 the proposed approach is more robust than expected from results shown in Fig. 4 in tracking jitter chnages. The analysis confirmed however that a significant portion of variance may be related to estimation error. This result should be taken into account when analyzing real data. From the results here shown, no strong significant differences were highlighted between the SWIPE' based and SIFT based approaches as regards jitter estimation.

The preliminary results on real dataset seem to indicate a significant jitter difference in vowel scored as anger. The proposed approach could be applied only using overlapping windows given the short duration of extracted vowels. Moreover the estimation of jitter values for short duration, low pitch vowels may not be possible or it may results in bad estimates given the small number of pitch periods available. An analysis of the time profile of pitch and jitter was out of the scope of this work, that was motivated by the need of characterizing the approach on short, quasi stationary vowels. Future work should take into account the analysis of long sentences. However the interpretation of the results on real datasets may take into account these considerations.

V. Conclusion

The proposed approach allows estimating pitch with good performances. Simulated data results show that an index proportional to jitter value can be estimated as well, allowing to employ this method for classification purposes. Preliminary results on real dataset indicate the potential application to running speech albeit with some limitations in the case of short utterances at low pitch.

References

[1] Bulut, M, Narayanan, S. On the robustness of overall F0-only modifications to the perception of emotions in speech. J Acoust Soc Am 2008;123(6):4547-4558

[2] Tao, J, Kang, Y, Li, A. Prosody Conversion From Neutral Speech to Emotional Speech. IEEE Trans Audio Speech Lang Processing 2006;14(4):1145-1154

[3] Vasilakis, M, Stylianou, Y. Spectral jitter modeling and estimation. Biomedical Signal Processing and Control 2009; Available on-line

[4] Boyanov B, Hadjitodorov S. Acoustic analysis of pathological voices. IEEE Eng Med Biol Mag 1997;16(4):74–82

[5] Camacho, A, Harris, J. A sawtooth wave form inspired speech estimatore for speech and music. J Acoust Soc Am 2008; 124(3):1638–1652

[6] Markel, JD, The SIFT algorithm for fundamental frequency estimation, IEEE Trans. Audio Electroacoust 1972; 20: 367-377

[7] Burkhardt, F, Paeschke, A, Rolfes, M, Sendlmeier, W, Weiss, B. A Database of German Emotional Speech, Proc. Interspeech, Lisbon, Portugal, 2005; 1517–1520