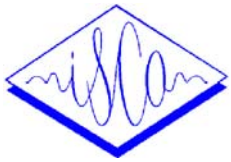

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

New Harmonicity Measures for Pitch Estimation and Voice Activity Detection

An-Tze Yu and Hsiao-chuan Wang

Department of Electrical Engineering
National Tsing Hua University, Hsinchu, Taiwan, ROC
yuat@ms25.hinet.net, hcwang@ee.nthu.edu.tw

Abstract

Harmonic structure can be easily recognized in the time-frequency representation of speech signals even in the diverse environment. The harmonicity is a measure of the completeness of harmonic structure. This paper extends the use of conventional harmonicity measure to the tasks of pitch estimation and voice activity detection. A set of hierarchical harmonicities, including grid, temporal, spectral and segmental harmonicities, is derived for this purpose. A series of experiments are conducted to show the effectiveness of using harmonicities in speech processing.

1. Introduction

The human's auditory system can easily perceive the existence of speech in very noisy environments. This comes from the auditory system of human being and the characteristics of speech signals. Many works have been done to investigate the functions of auditory system and apply them to speech processing tasks [1-2]. This paper focuses on the harmonic structure [3] of speech signals, which is the primary difference between the human speech and other environmental noises.

The harmonicity denotes the degree of a signal to be harmonic or periodic. For speech signal, the harmonicity is the synonym of voicing degree. Many methods have been developed to evaluate the harmonic structure of speech signal. The evaluation of R1/R0 is a simple but effective method. The harmonic-to-noise ratio (HNR) measured in the autocorrelation domain is another useful method [4]. Some cepstrum-based techniques are used to provide more efficient algorithms in determining the HNR [5-7]. The HNR can also be implemented in instantaneous frequency amplitude spectrum (IFAS) [8]. This paper extends the conventional harmonicity measure to a set of hierarchical harmonicities, including grid, temporal, spectral and segmental harmonicities. Through this extension, a systematic analysis of harmonic structure can be performed.

The set of harmonicities can be applied to many tasks of speech processing. This paper demonstrates the application of the harmonicities to robust pitch estimation and voice activity detection. The experimental results show the effectiveness of the proposed method.

2. New Harmonic Structure Measures

To provide a systematic harmonic structure analysis, a set of hierarchical harmonicities are proposed.

2.1. Grid harmonicity

The grid harmonicity measures the energy ratio of a harmonics to its surrounding noise. It involves three factors: (1) the local spectral dominance, (2) the temporal correlation, and (3) the harmonic spectral correlation. The local spectral dominance for m-th harmonics of a signal evaluated at frame n is defined as below,

$$h_D(n,m) = \frac{\sum_{k=(m-\eta)k_0}^{(m+\eta)k_0} S(n,k)\phi(k-mk_0)}{\sum_{k=(m-\eta)k_0}^{(m+\eta)k_0} S(n,k)\left(1-\phi(k-mk_0)\right)}, \quad (1)$$

where $S(n,k)$ is the magnitude spectrum density of the signal. k represents the frequency bin index,. k_0 is the frequency bin index of fundamental frequency. η is set to 0.5. $\phi(k)$ is a harmonics selector defined as

$$\phi(k) = e^{-\frac{k^2}{2\sigma^2}}, \quad (2)$$

where σ controls the width of harmonics selector.

The temporal variation of harmonics is measured by,

$$D_T(n,m) = \frac{\left|S(n,mk_0)-S(n-1,mk_0)\right|}{S(n,mk_0)+S(n-1,mk_0)} + \frac{\left|S(n+1,mk_0)-S(n,mk_0)\right|}{S(n+1,mk_0)+S(n,mk_0)}. \quad (3)$$

The temporal correlation is obtained from,

$$h_{tc}(n,m) = \frac{1}{1+e^{a(D_T(n,m)-b)}}, \quad (4)$$

where a and b control the property of transformation and are empirically set to 5 and 0.3, respectively. They are insensitive to testing data.

Harmonic correlation is measured by referencing its neighborhood local spectral dominances,

$$h_{hc}(n,m) = h_D(n,m-1) + h_D(n,m+1). \quad (5)$$

Then the grid harmonicity is expressed as

$$h(n,m) = h_D(n,m)h_{Tc}(n,m)h_{hc}(n,m), \quad (6)$$

2.2. Temporal harmonicity

Temporal harmonicity represents the strength of harmonicity of speech in a frame. It is the synonym of voicing degree and derived by summing the grid harmonicities over all harmonics,

$$H_T(n) = \sum_{m=1}^{M(n)} h(n,m)w_h(n,m), \quad (7)$$

where $M(n)$ is the number of harmonics at frame n. $w_h(n,m)$ is a weighting factor and expressed as,

$$w_h(n,m) = \frac{e(n,m)}{E_T(n)}, \quad (10)$$

where $e(n,m)$ and $E_T(n)$ are grid and temporal energies, respectively. They are computed by

$$e(n,m) = \sum_{k=(m-\eta)k_0}^{(m+\eta)k_0} S(n,k)\phi(k-mk_0), \quad (8)$$

and

$$E_T(n) = \sum_{m=1}^{M(n)} e(n,m), \quad (9)$$

2.3. Spectral harmonicity

The spectral harmonicity evaluates the integrity of harmonics in a speech segment. It is computed by summing the grid harmonicities over the specified segment,

$$H_H(s,m) = \sum_{n=SegBegin(s)}^{SegEnd(s)} h(n,m)w_t(n,m), \quad (11)$$

where SegBegin(s) and SegEnd(s) represent the first and last frame indexes belonging to segment s, respectively. $w_t(n,m)$ is a weighting factor defined as

$$w_t(n,m) = \frac{e(n,m)}{E_H(s,m)}, \quad (12)$$

where $E_H(s,m)$ is computed by.

$$E_H(s,m) = \sum_{n=SegBenin(s)}^{SegEnd(s)} e(n,m), \quad (13)$$

2.4. Segmental (phoneme) harmonicity

It is a whole evaluation for harmonic structure within a segment. The segmental or phoneme harmonicity is proposed and calculated as follows

$$H(s) = \sum_{n=SegBegin(s)}^{SegEnd(s)} H_T(n)W_T(s,n) = \sum_{m=1}^{M(s)} H_H(s,m)W_H(s,m), \quad (14)$$

where $M(s)$ is the number of harmonics at segment s. $W_T(s,n)$ and $W_H(s,m)$ are the weighting factors for temporal and spectral harmonicities, respectively. They are computed by

$$W_T(s,n) = \frac{E_T(n)}{E_S(s)}, \quad (15)$$

and

$$W_H(s,m) = \frac{E_H(s,m)}{E_S(s)}, \quad (16)$$

where $E_S(s)$ is the segmental energy defined as ,

$$E_S(s) = \sum_{n-SegBegin(s)}^{SegEnd(s)} E_T(n) = \sum_{m=1}^{M(s)} E_H(s,m), \quad (17)$$

3. Robust Pitch Estimation

The pitch estimation is necessary for several applications, including speech coding, speech recognition, speech enhancement, and prosody extraction. With such a wide range of interests, many pitch estimation algorithms had been developed to fit their specific applications [9]. Though pitch estimation has been developed for many years, it is difficult to robustly identify a correct pitch from its double or half values. The uncertainty in picking pitch candidate makes the pitch estimation algorithms complicated and not robust. Nevertheless, validating the pitch candidates by harmonic structure analysis can ease the problem. This is because a harmonic structure with correct pitch produces the highest harmonicity.

Experiments to investigate the performance of pitch estimation are conducted as follows: (1) Each pitch candidate is individually obtained from five domains, including the autocorrelation, the AMDF, the harmonic peaks, the sub-harmonic sum, and the cepstrum. (2) Pitch candidates with maximum likelihood at each frame are picked. (3) A 3-point median filter is used to filter out the outliers so that the final pitch track is obtained. (4) The baseline system does not apply the pitch candidate validation. (5) Two measures, the gross error and the normalized mean square error, are used to evaluate the performance [10].

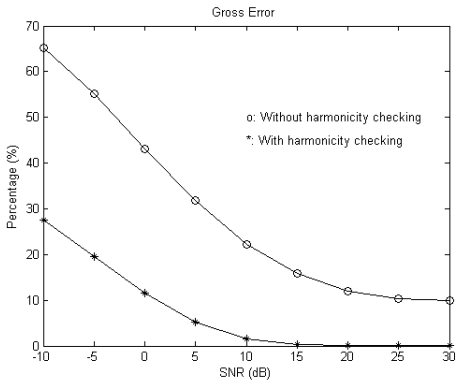

Figure 1. Average gross error rates of pitch estimation

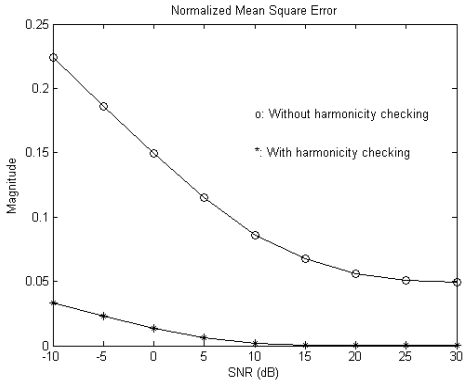

Figure 2. Average normalized mean square errors of pitch estimation

Ten utterances of Mandarin speech are used in the experiment. Nine kinds of noises listed in Table 1 are used as additive noises. The speech signals sampled at 8 kHz are degraded by the noises with specific SNR ranging from -5 to 30 dB, with a 5 dB step. The performances of pitch estimation from five domains are averaged and shown in Figures 1 and 2. These figures clearly show that the performance of pitch estimation can be greatly improved when the harmonicities are applied in the validation process.

4. Voice Activity Detection

The Voice Activity Detection (VAD) is one of the essential issues addressed in several speech applications. It is either advantageous or absolutely necessary to automatically distinguish speech and non-speech segments in the early stages of signal processing. The earlier algorithms are based on cepstral features or energy levels. Latest researches studied the fusion of conventional techniques and statistic approaches to improve the robustness of VAD [11]. The harmonicity provides another way to build a robust VAD.

Besides human speech, only a few sounds, like music, own harmonic structures. Hence, the existence of harmonic structure becomes a simple and robust indicator to identify a human speech. Table 1 lists the average temporal harmonicities of nine kinds of noises, and the average segmental harmonicity of speech signal. The big difference in harmonicities between noise and speech implies that the harmonicity is an appropriate parameter for speech/non-speech classification. Since unvoiced speech does not show the harmonic structure also, the voiced/unvoiced decision can be easily made by using harmonicity. Figure 3 displays the average segmental harmonicity of speech corrupted with several noises evaluated at various SNR conditions. The figure reveals that the harmonicity is a robust parameter for speech/non-speech or voiced/unvoiced speech detection.

Table 1. The average temporal harmonicities of noises

Noise type	Average harmonicity
Airport	1.72
Babble	1.75
Car	1.18
Exhibition	1.47
Restaurant	1.55
Street	1.56
Subway	1.22
Train	1.52
White	1.10
Human speech	5.13

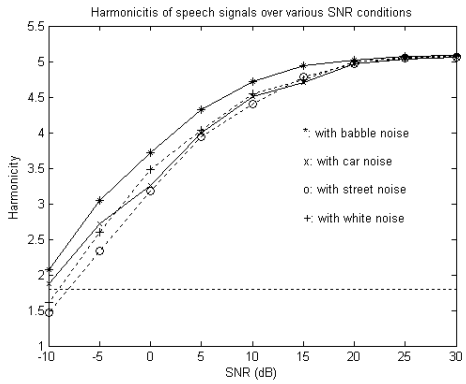

Figure 3. Average segmental harmonicity evaluated at various conditions. The dotted line is the maximal value of noises' harmonicity listed in table 1.

The experimental condition for using harmonicity to determine the voicing status of speech frame is the same as

that in pitch estimation experiment. The experimental result is plotted in Figure 4. The error rates of conventional voicing detection method based on harmonic-to-noise ratio (HNR) are plotted as the reference. The Voicing errors are defined as the percentage of frames such that the tracker and the reference disagree in voicing decision. The results reveal that the method based on hierarchical harmonicity outperforms HNR method.

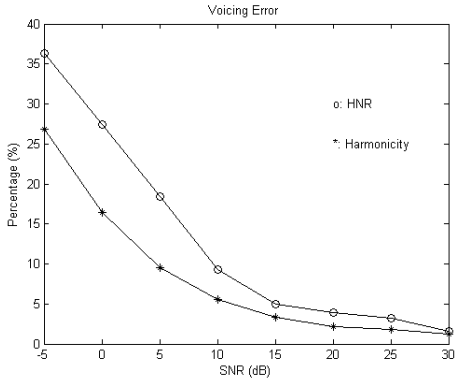

Figure 4. Average voicing error rates

5. Conclusion

This paper extends the conventional harmonicity to a set of hierarchical harmonicities, comprising grid, temporal, spectral and segmental harmonicities. Through this extension, detailed and systematic analysis of harmonic structure can be performed. The set of harmonicities can be applied to many speech processing tasks. The applications of harmonicity measures to pitch estimation and voice activity detection are demonstrated.

6. Acknowledgement

This research was partially sponsored by the National Science Council, Taiwan, under contract number NSC-92-2213-E-007-036.

7. References

[1] Hynek Hermansky, "Auditory Modeling in Automatic Recognition of Speech," in *Proc. Keele Workshop*, Keele, Sweden, 1996.

[2] Piero Cosi, "On The Use of Auditory Models in Speech Technology", in V. Roberto Ed., "*Lecture Notes in Artificial Intelligence: Intelligent Perception Systems*", Springer Verlag Publisher, *Vol. 745,* 1993.

[3] A.T. Yu and H.C. Wang, "New speech harmonic structure measure and it application to post speech enhancement," To appear in *Proc.ICASSP 2004*.

[4] Paul Boersm, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," IFA Proceedings 17, 1993

[5] de Krom, G., "A cepstrum based technique for determining an harmonics-to-noise ratio in speech signals," J. Speech Hear. Res. 36(2), 254-266, 1993

[6] Qi, Y. and Hillman, RE., Temporal and spectral estimations of harmonics-to-noise ratio human voice signals. J. Acoust. Soc. Am. 102(1), 537-543, 1997

[7] Murphy, Peter J., "A cepstrum-based harmonics-to-noise ratio in voice signals", *Proc. ICSLP* 2000, vol. 4, 672-675.

[8] Dhany Arifianto, Takao Kobayashi, "IFAS-based Voiced/Unvoiced Classification of Speech Signal,'" *Proc. the 2003 IEEE International Conference on Acoustics, Speech and Signal Processing,* ICASSP 2003, vol.I, pp.812-815, 2003.

[9] Alain de Cheveign´e and Hideki Kawahara, "Comparative evaluation of *F*0 estimation algorithms," *Proc. Eurospeech 2001.*

[10] Kavita Kasi and Stephen A. Zahorian,"Yet another algorithm for pitch tracking," *Proc. ICASSP* 2002.

[11] S. Gökhun Tanyer and Hamza Özer," Voice Activity Detection in Nonstationary Noise," *IEEE Transactions on Speech and Audio Processing,* Vol. 8, no. 4, July 2000