Short Article

A Measure of Smoothness in Synthesized Speech

Phung Trung Nghia[1], **Phung Thi Thu Hien**[2], **Nguyen Van Tao**[1],
Pham Thi Mai Huong[1], **Nguyen Thi Bich Diep**[1]

[1] Thai Nguyen University of Technology, Thai Nguyen, Vietnam
[2] Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Vietnam

Correspondence: Phung Trung Nghia, ptnghia@ictu.edu.vn

Communication: received 22 June 2015, revised 29 September 2015, accepted 21 March 2016
Online publication: 16 June 2016, Digital Object Identifier: 10.21553/rev-jec.106
The associate editor coordinating the review of this article and recommending it for publication was Dr. Le Vu Ha.

Abstract– **The articulators typically move smoothly during speech production. Therefore, speech features of natural speech are generally smooth. However, over-smoothness causes "muffleness" and, hence, reduction in ability to identify emotions/expressions/styles in synthesized speech that can affect the perception of naturalness in synthesized speech. In the literature, statistical variances of static spectral features have been used as a measure of smoothness in synthesized speech but they are not sufficient enough. This paper proposes another measure of smoothness that can be efficiently applied to evaluate the smoothness of synthesized speech. Experiments showed that the proposed measure is reliable and efficient to measure the smoothness of different kinds of synthesized speech.**

Keywords– **speech synthesis, speech quality, speech smoothness measure, global variance.**

1 Introduction

Although the definition of smoothness in speech has not been clearly mentioned, it is implicit in many studies [1–7], in which speech smoothness can be considered as a result of transitions in speech. The slower transitions cause the speech smoother and, vice-versa, the more rapid transitions cause the rougher speech.

The articulators typically move slowly during speech production [1]. Therefore, speech features of natural speech are generally smooth. The over-roughness can reduce the naturalness of synthesized speech. However, rapid changes in speech features naturally occur in some cases such as in plosives [1]. These are natural over-roughness or natural discontinuities in speech features. Besides natural over-roughness, several kinds of unexpected over-roughness in speech, such as discontinuities caused by mismatch-context errors in speech synthesis or by noisy recording environments, can reduce the naturalness of synthesized speech [2], or of recorded speech [3].

On the contrary, the over-smoothness of synthesized speech [4] also reduces the naturalness of synthesized speech due to several reasons presented below. Over-smoothness causes "muffleness" in synthesized speech [5] that affects its naturalness. Degree of articulation (DoA) provides information about style and/personality [6]. It is characterized by modifications of the phonetic context, the speech rate, and spectral dynamics (vocal tract rate of change). Over-smooth speech with too-slow transitions may affect the production of the appropriate DoA and the important information about style/personality may be lost. Over-smoothness may be acceptable for reading speech or neutral speech but not suitable for expressive speech. Mainly, the range and the velocity of the tongue tip movements are the primary modulation parameters associated with emotional expression [7]. Therefore, too-smooth speech with slow movements cannot be efficient to represent some kinds of emotional speech with high movements of the tongue tip. Lieberman and Michaels [8] found that smoothing the F_0 contours could reduce the recognition rate of the emotion of speech. The smoother speech, the lower recognition rate of the emotion of speech. Over-smoothness can eliminate F_0 and spectral fluctuations that are important in singing voice synthesis and perception and in expressive speech synthesis and perception [9]. While linguistic information in speech is critical for its intelligibility, non-linguistic information in speech (e.g., emotion, expression, individuality) is important to perceive its naturalness. Therefore, over-smoothness reduces the naturalness of synthesized speech.

Consequently, both over-smoothness and over-roughness can reduce the naturalness of synthesized speech. Therefore, instead of synthesizing too-smooth or too-rough speech, the optimal smoothness that naturally exists in the original speech has to be reached to ensure the naturalness of synthesized speech.

In this paper, a speech smoothness measure will be proposed and a concept of "appropriate smoothness" in speech will be defined as the smoothness that approximates the optimal smoothness naturally existed in the original human speech. With appropriate smoothness, speech is supposed to be natural. This appropriate smoothness depends on the content of speech and is different between vowels and consonants. It also depends on the observed speech features.

2 Measuring Speech Smoothness Using the Global Variance

In the literature, statistical variances of static spectral features have been widely used as measures of smoothness of synthesized speech [5] or noisy speech [10]. By generating a speech that has a global variance (GV) close to that of the original speech, the synthesized speech is expected to be natural [5]. On the other hand, synthesized speech is expected to have an "appropriate smoothness" if it has GVs of static spectral features close to those of the original speech. By minimizing the variance in smoothed signal spectral density, several smoothing methods for noisy speech enhancement were proposed in [10]. The enhanced speech was smoother and its quality was significantly enhanced.

Time-domain and spectral-domain GVs of a static spectral feature are defined as

$$\mathrm{GV}_t = \frac{1}{P}\sqrt{\sum_{p=1}^{P}(\mathrm{var}_t(p))^2}, \quad (1)$$

$$\mathrm{GV}_s = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(\mathrm{var}_s(i))^2}, \quad (2)$$

where var_t and var_s are the variances in the time domain and the spectral domain of the spectral feature, P is the dimension of the feature, N is the length in the time domain of the feature.

The distance between the GVs of the synthesized speech and the original speech can be used to measure the closeness between the smoothness of synthesized speech and that of the original speech. Here, the distances between the GV of a synthesized static feature and the GV of a static feature of the original speech (GV^*) are given by

$$\mathrm{DGV}_t(x) = \frac{\mathrm{GV}_t^* - \mathrm{GV}_t(x)}{\mathrm{GV}_t^*}, \quad (3)$$

$$\mathrm{DGV}_s(x) = \frac{\mathrm{GV}_s^* - \mathrm{GV}_s(x)}{\mathrm{GV}_s^*}. \quad (4)$$

Although variances of static spectral features can measure smoothness in speech features, they do not exactly represent smoothness of speech features. Figure 1 gives an example of two signals that have the same mean and variance but different smoothness. Figure 2 gives an example of an LSF sequence synthesized by a Hidden Markov Model (HMM) with and without the GV. Two drawbacks of the GV of static spectral features on measuring smoothness of features are shown in this figure. First, although increasing the GV of static spectral features can make synthesized speech closer to the original speech, the synthesized speech is still oversmooth compared to the original speech. Second, the estimated GV may be inaccurate due to some reasons such as the limitation of the training data.

Due to the limitation of statistical variances of static spectral features on measuring smoothness of speech features, it is important to propose a new and efficient speech smoothness measure in the field of speech synthesis.

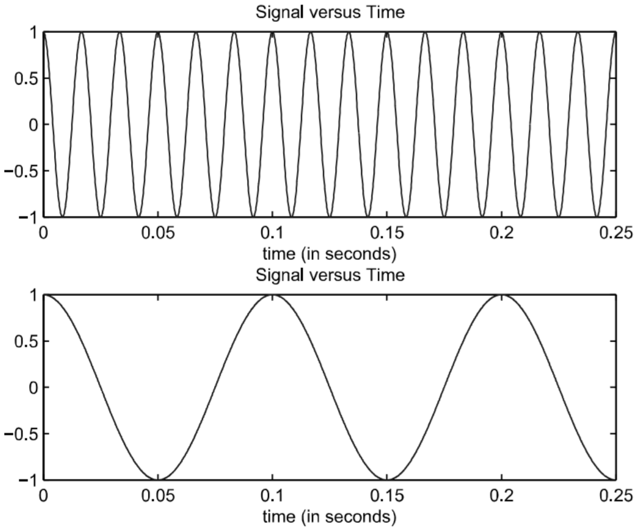

Figure 1. Two signals with the same mean and variance but different smoothness.

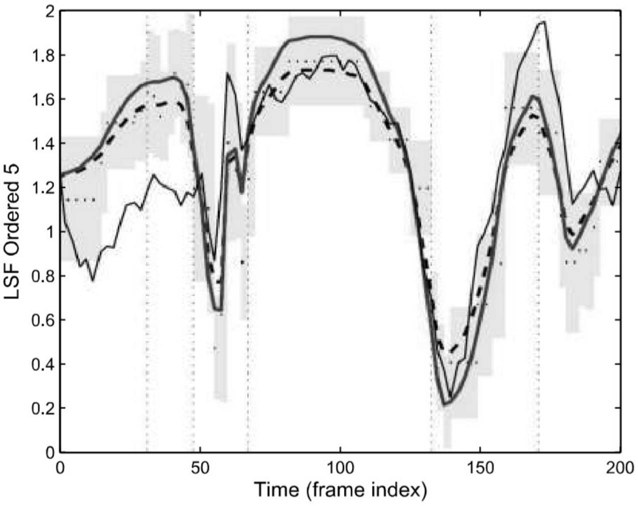

Figure 2. A LSF sequence synthesized by HMM with GV (bold curve), without GV (dashed curve), and the original one (thin curve): the shades in both sides of the sequence are the standard deviation.

3 The Proposed Speech Smoothness Measure

Smoothness in signals, time series is a result of transitions in the signals and time series. Researches on mathematics, time series analysis, and signal processing show that the smoothness of a curve, a time series, a signal, or a feature of a signal can be measured based on the "curvature" of the envelope of the function, the time series, the signal, or the feature [11]. The curvature is usually computed by using the second-order derivative of the curve, the time series or the signal [11]. Using second-order derivative can also represent the transition, or the changing rate in speech. Therefore, second-order derivative was used to define a measure of smoothness in speech.

In time series analysis and discrete signal processing, it has been revealed that delta of delta (delta-delta) can represent a simple second-order derivative [11]. Therefore, the square sum of the variance of the delta-

delta sequence was used to define the "Global Speech Smoothness Measure" (GSM) in this research, which is based on the formulation to measure a "global" smoothness for time series presented in [11].

Suppose that we need to measure the smoothness of a speech feature $\mathbf{X}_{P\times N}$

$$\mathbf{X}_{P\times N} = \{X_i^p\}, \quad p = 1,\ldots,P, \quad i = 1,\ldots,N, \quad (5)$$

where $P > 1$ corresponds to a spectral feature such as LSF and $P = 1$ corresponds to a prosodic feature such as the F_0 contour.

For one-dimensional features ($P = 1$) such as in F_0, smoothness can only be measured and observed in the time domain, named temporal smoothness in this paper. However, for P-dimensional features ($P > 1$) such as in LSF, smoothness can be measured and observed in both time and spectral domain, named temporal and spectral smoothness in this paper.

3.1 Temporal Speech Smoothness Measure

Delta in the time domain for spectral sequence p is expressed as

$$\Delta t(\mathbf{X})_{P\times(N-1)} = \{X_{i+1}^p - X_i^p\}, \quad (6)$$

with $p = 1,\ldots,P$, and $i = 1,\ldots,N-1$.

Its delta-delta in the time domain is expressed as

$$\Delta^2 t(\mathbf{X})_{P\times(N-2)} = \{\Delta t_{i+1}^p - \Delta t_i^p\}, \quad (7)$$

with $p = 1,\ldots,P$, and $i = 1,\ldots,N-2$.

The variance of delta-delta in the time domain for spectral sequence p is given by

$$\mathrm{var}\{\Delta^2 t(p)\} = \frac{1}{N-2}\sum_{i=1}^{N-2}\left(\Delta^2 t_i^p - \overline{\Delta^2 t_p}\right)^2, \quad (8)$$

where

$$\overline{\Delta^2 t_p} = \frac{1}{N-2}\sum_{i=1}^{N-2}\Delta^2 t_i^p. \quad (9)$$

Finally, the GSM in the time domain (temporal GSM) is defined as

$$\mathrm{GSM}_t = \frac{1}{P}\sqrt{\sum_{p=1}^{P}\left(\mathrm{var}\{\Delta^2 t(p)\}\right)^2}. \quad (10)$$

3.2 Spectral Speech Smoothness Measure

Delta in the spectral domain for frame i is expressed as

$$\Delta s(\mathbf{X})_{(P-1)\times N} = \{X_i^{p+1} - X_i^p\}, \quad (11)$$

with $p = 1,\ldots,P-1$, and $i = 1,\ldots,N$.

Its delta-delta in the spectral domain is expressed as

$$\Delta^2 s(\mathbf{X})_{(P-2)\times N} = \{\Delta s_i^{p+1} - \Delta s_i^p\}, \quad (12)$$

with $p = 1,\ldots,P-2$, and $i = 1,\ldots,N$.

The variance of delta-delta in the spectral domain for frame i is given by

$$\mathrm{var}\{\Delta^2 s(i)\} = \frac{1}{P-2}\sum_{p=1}^{P-2}(\Delta^2 s_i^p - \overline{\Delta^2 s_i})^2, \quad (13)$$

where

$$\overline{\Delta^2 s_i} = \frac{1}{P-2}\sum_{p=1}^{P-2}\Delta^2 s_i^p. \quad (14)$$

Finally, the GSM in the spectral domain (spectral GSM) is defined as

$$\mathrm{GSM}_s = \frac{1}{N}\sqrt{\sum_{i=1}^{N}\left(\mathrm{var}\{\Delta^2 s(i)\}\right)^2}. \quad (15)$$

Based on the definitions of temporal and spectral GSMs (GSM_t, GSM_s) in Equations (10) and (15), it reveals that the smaller the GSM, the smoother the speech feature, while the larger the GSM, the rougher the speech feature. As a result, GSM can be used to measure the smoothness of a speech feature. However, instead of synthesizing too-smooth or too-rough speech, an appropriate smoothness has to be reached to ensure naturalness of synthesized speech. Therefore, instead of directly using the GSM, the distance between the GSM of synthesized speech and that of the corresponding original speech (GSM*), is proposed to measure smoothness of synthesized speech.

3.3 Distance of Global Smoothness Measure

Distance of GSM (DGSM) in the time domain and the spectral domain are defined as below, respectively:

$$\mathrm{DGSM}_t(x) = \frac{\mathrm{GSM}_t^* - \mathrm{GSM}_t(x)}{\mathrm{GSM}_t^*}, \quad (16)$$

$$\mathrm{DGSM}_s(x) = \frac{\mathrm{GSM}_s^* - \mathrm{GSM}_s(x)}{\mathrm{GSM}_s^*}. \quad (17)$$

Based on these definitions, the following are revealed:

- If DGSM is positive, the synthesized speech is smoother than the original speech.
- If DGSM is negative, the synthesized speech is rougher than the original speech.
- If the absolute of DGSM is close to zero, the synthesized speech has an "appropriate smoothness".
- If DGSM is positive and its absolute is large, the synthesized speech is over-smooth.
- If DGSM is negative and its absolute is large, the synthesized speech is over-rough.

In the next section, DGSM will be used to measure the smoothness of speech synthesized by some popular speech coders and synthesizers, to confirm the reliability of this proposed measure.

4 Experiments

This section presents experiments of using DGV and DGSM to measure smoothness of synthesized speech.

Experiment 1: 100 utterances extracted from Vietnamese speech corpus named DEMEN567 [12] were analyzed/synthesized by a high-quality speech coder named STRAIGHT [13] and synthesized with a HMM speech synthesizer with GV [14].

Experiment 2: 100 utterances extracted from English speech corpus named TIMIT [15] were analyzed/synthesized by STRAIGHT and synthesized with a HMM speech synthesizer with GV [5].

Table I
DGV in Time Domain of Vietnamese Speech

DGV	STRAIGHT	HMM with GV
Mean	-0.007	-0.016
95% Confidence	0.016	0.038

Table II
DGV in Spectral Domain of Vietnamese Speech

DGV	STRAIGHT	HMM with GV
Mean	0.01	-0.019
95% Confidence	0.0007	0.007

Table III
DGSM in Time Domain of Vietnamese Speech

DGSM	STRAIGHT	HMM with GV
Mean	-0.0809	0.5534
95% Confidence	0.01	0.0232

Table IV
DGSM in Spectral Domain of Vietnamese Speech

DGSM	STRAIGHT	HMM with GV
Mean	-0.098	0.1685
95% Confidence	0.018	0.0218

Table V
DGV in Time Domain of English Speech

DGV	STRAIGHT	HMM with GV
Mean	0.014	-0.008
95% Confidence	0.009	0.012

Table VI
DGV in Spectral Domain of English Speech

DGV	STRAIGHT	HMM with GV
Mean	-0.011	-0.021
95% Confidence	0.002	0.013

Table VII
DGSM in Time Domain of English Speech

DGSM	STRAIGHT	HMM with GV
Mean	-0.0114	0.316
95% Confidence	0.02	0.017

Table VIII
DGSM in Spectral Domain of English Speech

DGSM	STRAIGHT	HMM with GV
Mean	-0.059	0.253
95% Confidence	0.021	0.016

Results of experiment 1 are shown in Tables I-IV. Results of experiment 2 are shown in Tables V-VIII. The results in Tables I, II (for Vietnamese speech), V, VI (for English speech) show that DGV of speech synthesized by HMM in both time and spectral domains are very small and are almost equivalent with speech analyzed/synthesized by STRAIGHT. However, speech synthesized by HMM with GV is still over-smooth in the time domain as shown in Figure 2. Additionally, subjective evaluations in many researches show that speech synthesized by HMM is still muffled and over-smooth [14]. *Therefore, these results support that DGV has bad correlations with perceived quality of synthesized speech.*

The results in Tables III, IV (for Vietnamese speech), VII, VIII (for English speech) show that DGSM of speech analyzed/synthesized by STRAIGHT is very small. These results are close to the fact that speech analyzed/synthesized by the high-quality speech coder STRAIGHT is very close to the original speech in term of smoothness. DGSM of speech synthesized by HMM is positive and its absolute is very large. Then, speech synthesized by the HMM synthesizer is shown to be over-smooth in both time and spectral domains when measuring by DGSM. These conclusions are identical as theoretical and experimental results of previous studies [5, 12, 15]. *Therefore, they support that the proposed DGSM has good correlations with perceived quality of synthesized speech.*

5 Conclusion

This paper has considered the need of a speech smoothness measure to evaluate methods of synthesizing speech. The disadvantages of using GV to measure speech smoothness was described. Then, using DGSM, which exploits the square sum of the variance of the delta-delta sequence, to measure speech smoothness was presented and discussed. Experiments comparing DGV and DGSM have showed that, while DGV is not reliable, DGSM is reliable and efficient to measure smoothness of different kinds of synthesized speech.

References

[1] G. Fant, *Acoustic theory of speech production*. The Netherlands: Mouton-The Hague, 1960.

[2] K. K. Paliwal, "Interpolation properties of linear prediction parametric representations," in *Fourth European Conference on Speech Communication and Technology*, 1995, pp. 1016–1032.

[3] R. McArdle and R. H. Wilson, "Speech perception in noise: The basics," *SIG 6 Perspectives on Hearing and Hearing Disorders: Research and Diagnostics*, vol. 13, no. 1, pp. 4–13, 2009.

[4] Y.-Y. Chen, T.-W. Kuan, C.-Y. Tsai, J.-F. Wang, and C.-H. Chang, "Speech variability compensation for expressive speech synthesis," in *International Conference on Orange Technologies (ICOT)*. IEEE, 2013, pp. 210–213.

[5] T. Tomoki and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.

[6] G. Beller, N. Obin, and X. Rodet, "Articulation degree as a prosodic dimension of expressive speech," in *Fourth International Conference on Speech Prosody*, Campinas, Brazil, 2008, pp. 677–681.

[7] S. Lee, E. Bresch, and S. Narayanan, "An exploratory study of emotional speech production using functional data analysis techniques," in *Proceedings of the 7th International Seminar on Speech Production*, Ubatuba, Brazil, 2006, pp. 525–532.

[8] P. Lieberman and S. B. Michaels, "Some aspects of fundamental frequency and envelope amplitude as related

to the emotional content of speech," *The Journal of the Acoustical Society of America*, vol. 34, no. 7, pp. 922–927, 1962.

[9] S. W. Lee, S. T. Ang, M. Dong, and H. Li, "Generalized F0 modelling with absolute and relative pitch features for singing voice synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 429–432.

[10] R. Martin, "Spectral Subtraction Based on Minimum Statistics," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 1994, pp. 1182–1185.

[11] E. B. Dagum and M. Morry, "Basic issues on the seasonal adjustment of the Canadian consumer price index," *Journal of Business & Economic Statistics*, vol. 2, no. 3, pp. 250–259, 1984.

[12] L. C. Mai and D. N. Duc, "Design of Vietnamese speech corpus and current status," in *Proceedings of the International Symposium on Chinese Spoken Language Processing (ISCSLP)*, vol. 6, 2006, pp. 748–758.

[13] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology*, vol. 27, no. 6, pp. 349–353, 2006.

[14] T. T. Vu, M. C. Luong, and S. Nakamura, "An HMM-based Vietnamese speech synthesis system," in *2009 Oriental COCOSDA International Conference on Speech Database and Assessments*. IEEE, 2009, pp. 116–121.

[15] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.