EIGENRESIDUALS FOR IMPROVED PARAMETRIC SPEECH SYNTHESIS

Thomas Drugman [1], *Geoffrey Wilfart* [2], *Thierry Dutoit* [1]

[1] TCTS Lab, Faculté Polytechnique de Mons - 31, Boulevard Dolez, 7000, Mons, Belgium
[2] Research and Development, Acapela Group - 33, Boulevard Dolez, 7000, Mons, Belgium
phone: + (32) 65 37 47 49, fax: + (32) 65 37 47 29
email: thomas.drugman@fpms.ac.be

ABSTRACT

Statistical parametric speech synthesizers have recently shown their ability to produce natural-sounding and flexible voices. Unfortunately the delivered quality suffers from a typical *buzziness* due to the fact that speech is vocoded. This paper proposes a new excitation model in order to reduce this undesirable effect. This model is based on the decomposition of pitch-synchronous residual frames on an orthonormal basis obtained by Principal Component Analysis. This basis contains a limited number of *eigenresiduals* and is computed on a relatively small speech database. A stream of PCA-based coefficients is added to our HMM-based synthesizer and allows to generate the voiced excitation during the synthesis. An improvement compared to the traditional excitation is reported while the synthesis engine footprint remains under about 1Mb.

1. INTRODUCTION

For the last decade, Unit Selection-based methods [1] have clearly emerged in speech synthesis. These techniques rely on a huge corpus (typically several hundreds of Mb) covering as much as possible the diversity one can find in the speech signal. During synthesis, speech is obtained by concatenating natural units picked up from the corpus. As the database contains several examples for each speech unit, the problem consists in finding the best path through a lattice of potential candidates by minimizing a selection and concatenation cost. This approach generally generates speech with high naturalness and intelligibility. However quality may degrade severely when an under-represented unit is required or when a bad jointure (between two selected units) causes a discontinuity.

More recently, a new synthesis method has been proposed: the Statistical Parametric Speech Synthesis [2]. This approach relies on a statistical modeling of speech parameters. After a training step, it is expected that this modeling has the ability to generate realistic sequences of such parameters. The most famous technique derived from this framework is certainly the HMM-based speech synthesis [3], which obtained in recent subjective tests a performance comparable to Unit Selection-based systems [4]. An important advantage of such a technique is its flexibility for controling speech variations (such as emotions or expressivity) and for easily creating new voices (via statistical voice conversion). Its two main drawbacks, due to its inherent nature, are:

- the lack of naturalness of the generated trajectories. The statistical processing tends to remove details in the feature evolution; generated trajectories are oversmoothed, which makes the synthetic speech sound muffled. Some approaches considering global variance [5] or trajectory HMMs [6] have been proposed in order to reduce this detrimental effect.
- the "buzziness" of produced speech, which suffers from a typical vocoder quality.

While the parameters characterizing spectrum and prosody are rather well-established, improvement can be expected by adopting a more suited excitation modeling. Indeed the traditional excitation considers either a white noise or a pulse train during unvoiced or voiced segments respectively. Inspired from the physiological process of phonation where the glottal signal is composed of a combination of periodic and aperiodic components, the use of a Mixed Excitation (ME) has been proposed. The ME is generally achieved as in Figure 1. In [7], the filter coefficients were derived from bandpass voicing strenghts. In [8], state-dependent high-degree filters were directly trained using a closed loop procedure. The integration of a Liljencrants-Fant waveform as a modeling of the glottal source, possibly producing different voice qualities by varying the LF parameters, was proposed in [9]. In [10], we suggested the use of a codebook of typical pitch-synchronous residual frames. All these approaches reported a certain improvement with regard to the traditional excitation.

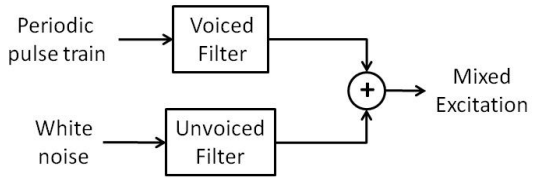

Figure 1: Mixed Excitation workflow.

This paper proposes a new excitation model in order to reduce the buzziness of parametric speech synthesizers. This model is based on the decomposition of pitch-synchronous residual frames on an orthonormal basis obtained by Principal Component Analysis (PCA). This basis contains a limited number of *eigenresiduals* and is computed on a relatively small speech database ($\approx$ 20 min), from which a dataset of voiced frames is extracted. These frames have the particularity of being centered on a Glottal Closure Instant (GCI), two-period long and Hanning-windowed. Furthermore they are resampled on a fixed number of points and normalized in energy. This is required to ensure inter-frame coherence before applying PCA. Once the PCA transform is calculated, the whole corpus is analyzed and PCA-based parameters are extracted. This allows us to enhance our HMM-based speech synthesizer with a new stream of excitation parameters, be-

sides the traditional pitch feature. During synthesis, voiced frames are derived from the generated PCA coefficients and overlap-added so as to obtain the excitation signal (Figure 2).

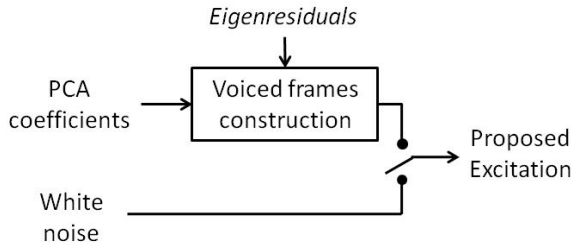

Figure 2: The proposed excitation workflow.

The paper is structured as follows. Section 2 describes the way the eigenresiduals are obtained from a speech database. For this, a dataset of normalized residual frames is extracted (2.1) and PCA is calculated on it, allowing dimensionality reduction (2.2). Section 3 details how this approach is integrated into an HMM-based speech synthesizer, relying on the framework available in [11], and presents the results of perceptual listening tests. Experiments throughout the paper were performed on three speakers: AWB (Scottish male) and SLT (US female) from the publicly available CMU ARCTIC speech database [12], and Bruno (French male) kindly provided by Acapela Group. Finally Section 4 concludes and proposes some guidelines for future works.

2. EIGENRESIDUALS THROUGH SPEECH ANALYSIS

The goal of this Section is to describe how eigenresiduals are obtained from a speech database. A dataset of normalized pitch-synchronous residual frames is first extracted (Section 2.1). As this set contains comparable data, Principal Component Analysis (PCA) can be calculated from it (Section 2.2) and a limited number of eigenresiduals can be retained.

2.1 Obtaining normalized pitch-synchronous residual frames

Mel-Generalized Cepstral coefficients (MGC) have been designed so as to accurately and robustly capture the spectral envelope of speech signals [13]. The workflow presented in Figure 3 thus performs MGC analysis with $\delta = 0.42$ ($Fs = 16kHz$) and $\gamma = -1/3$, as these values gave the best perceptual results in [4]. Residual signals are then obtained by inverse filtering. As previously mentioned, an important characteristic of our residual frames is that they are centered on Glottal Closure Instants (GCIs). In order to locate GCIs, we use a method based on the Center of Gravity (CoG) in energy of the speech signal, as suggested in [14]. While this gives straightforward results in most of cases, some residual segments are more contentious. Figure 4 exhibits how a peak-picking technique coupled with the detection of zero-crossings (from positive to negative) of the CoG signal removes possible ambiguities on the GCI positions.

Residuals are then windowed by a two-period Hanning window. To ensure a point of comparison between residual frames before applying PCA, GCI-alignment is not sufficient. Indeed frames probably come from different prosodic contexts, so that some normalization in both pitch and energy

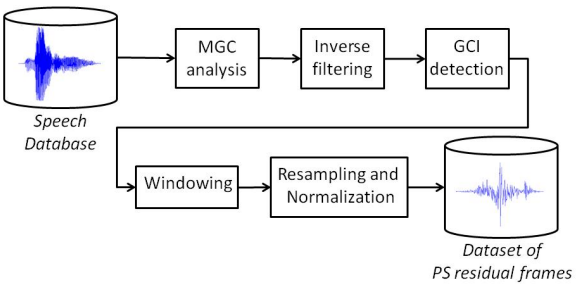

Figure 3: Obtaining a dataset of normalized pitch-synchronous residual frames from a speech database.

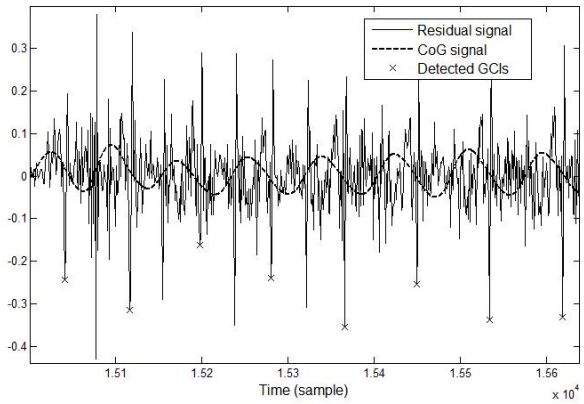

Figure 4: Determining the GCI location using the Center of Gravity technique.

is required. Pitch normalization is achieved by resampling, which retains the residual most important features. As a matter of fact, assuming that the residual obtained by inverse filtering approximates the glottal flow first derivative, resampling this signal preserves the *open quotient*, *asymmetry coefficient* (and consequently the F_g/F_0 ratio, where F_g stands for the *glottal formant* frequency) as well as the return phase characteristics (see [15]). Care has to be taken here when choosing the normalized pitch value, as this step will condition the synthesis quality. Indeed, at synthesis time, residual frames will be obtained by resampling a combination of eigenresiduals. If these have not a sufficiently low pitch, the ensuing upsampling will compress the spectrum and cause the appearance of "energy holes" at high frequencies. In order to avoid it, the speaker's pitch histogram $P(F_0)$ is analyzed and the normalized pitch value F_0^* we chose typically satisfies:

$$\int_{F_0^*}^{\infty} P(F_0)dF_0 \approx 0.8 \quad (1)$$

such that only 20% frames will be slightly upsampled at synthesis time (see Figure 5).

At this point, we have thus at our disposal a dataset of GCI-centered, Hanning-windowed, pitch and energy-normalized residual frames which is suited for applying PCA.

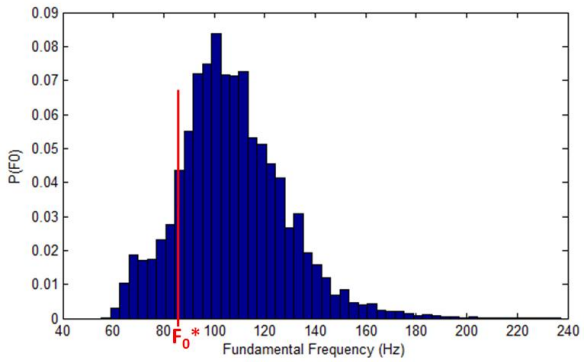

Figure 5: Analyzing a pitch histogram in order to determine the normalized pitch value (here for the male speaker Bruno).

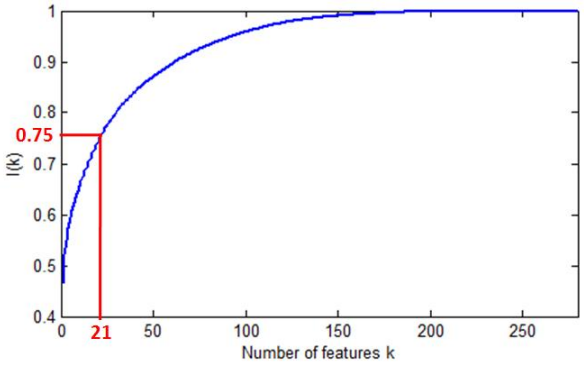

Figure 6: "Information" rate when using k eigenresiduals for speaker AWB.

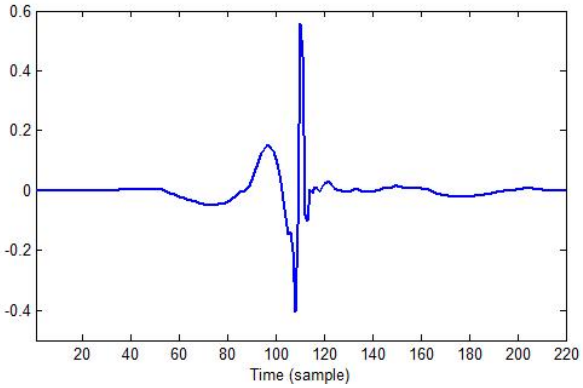

Figure 7: The first eigenresidual for the female speaker SLT.

2.2 Eigenresiduals computation

Principal Component Analysis (PCA) is an orthogonal linear transformation which applies a rotation of the axis system so as to obtain the best representation of the input data, in the Least Squared (LS) sense [16]. It can be shown that the LS criterion is equivalent to maximizing the data dispersion along the new axes. PCA can then be achieved by calculating the eigenvalues and eigenvectors of the data covariance matrix.

Let us assume that our dataset consists of N residual frames of m samples. PCA computation will lead to m eigenvalues λ_i with their corresponding eigenvectors μ_i (here called *eigenresiduals*). λ_i represents the data dispersion along axis μ_i and is consequently a measure of the "information" this eigenresidual conveys on the dataset. This is important in order to apply dimensionality reduction. Let us define $I(k)$, the "information rate" when using k eigenresiduals, as the ratio of the dispersion along these k axes over the total dispersion:

$$I(k) = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \quad (2)$$

Figure 6 displays this variable for the male speaker AWB ($m = 280$ in this case). Through subjective tests on an Analysis-Synthesis application, we observed that choosing k such that $I(k)$ is greater than about 0.75 has almost inaudible effects when compared to the original file. Back to the example of Figure 6, this implies that about 20 eigenresiduals can be efficiently used for this speaker.

To give an idea of what an eigenresidual looks like, Figure 7 exhibits the first eigenresidual, interpreted as the principal pattern arising from the data, for the female speaker SLT ($m = 220$). A strong similarity with glottal flow models can be noticed, mainly during the glottal open phase.

3. EIGENRESIDUALS AND PARAMETRIC SPEECH SYNTHESIS

3.1 Our HMM-based speech synthesizer

As previously mentioned, the HMM framework is very suitable in statistical parametric speech synthesis, as it has the ability to produce compact models. From a training perspective, it also uses well-known algorithms with guaranteed convergence. HMMs are built on the assumption that the speech signal is made of a concatenation of short quasi-stationary segments. Each segment corresponds to a state, and is viewed as a realisation of some statistical distribution, generally modeled by a Gaussian mixture. The training step may then look very close to building a speech recognizer, except that we use a context oriented clustering via binary decision trees.

Our feature vector consists of the 24-th order MGC parameters, log-F_0, and the PCA coefficients whose order has been determined as explained in Section 2.2, concatenated together with their first and second derivatives. Log-F_0 and PCA coefficients are however only meaningful in voiced segments of the signal. This leads to a modeling problem. On the one hand, some "value" needs to be provided for these parameters, even in unvoiced regions. On the other hand, providing a value in unvoiced regions will necessarily affect the Gaussian models.

In [17], authors propose an elegant solution and introduce the concept of Multi-Space Distribution (MSD). A MSD considers that the sample space is composed of several "spaces" that can have different dimensions (possibly 0). Each space has its own weight and, if the space dimension is non-zero, its probability distribution function (pdf). An event of dimension n then consists of a set of n-dimensional spaces it belongs to, and its probability in each of these spaces. As shown in [17], MSD is a general framework including standard discrete and continuous density pdfs. Authors also show that it is suitable in the context of HMM modeling by deriv-

ing reestimation formulae.

For our purpose, we chose to model each of log-F_0 and PCA coefficients by 2-space distributions. One of the spaces corresponds to the voiced regions and has a diagonal-covariance single-Gaussian distribution. The other space for unvoiced regions, has zero-dimensionality, and is characterized only by its weight. The same model is adopted for first and second derivatives of the parameters. For the sake of simplicity, we chose to model static features, first and second derivatives of our MSD parameters as independent streams. Finally, we use a single-space diagonal-covaraiance single-Gaussian distribution for MGCs and their first and second derivatives. Our model therefore uses 7 streams, defined as follows:

- MGCs + ΔMGCs + ΔΔMGCs,
- log-F_0,
- Δlog-F_0,
- ΔΔlog-F_0,
- PCAs,
- ΔPCAs,
- and ΔΔPCAs.

We use 5-state left-to-right context-dependent phoneme models, using pdfs described above. A state duration model is also determined from HMM state occupancy statistics [18]. During the speech synthesis process, the most likely state sequence is first determined according to the duration model. The most likely feature vector sequence associated to that state sequence is then generated, as described in [19]. Finally, these feature vectors are fed into a vocoder to produce the speech signal.

The vocoder workflow is depicted in Figure 8. The generated F_0 value commands the voiced/unvoiced decision. During unvoiced frames, white noise is used. On the opposite, the voiced frames are constructed according to the synthesized PCA coefficients. A first version is obtained by linear combination with the eigenresiduals extracted as detailed in Section 2. Since this version is size-normalized, a conversion towards the target pitch is required. As justified in Section 2.1, this can be achieved by resampling. The choice we made during the normalization of a sufficiently low pitch (cf. Equation 1) is now clearly understood as a constraint for avoiding the emergence of energy holes at high frequencies. Frames are then overlap-added so as to obtain the excitation signal. The so-called Mel Log Spectrum Approximation (MLSA) filter, based on the generated MGC coefficients, is finally used to get the synthesized speech signal.

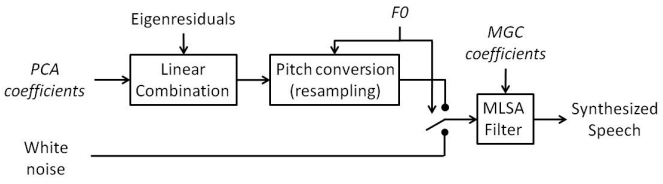

Figure 8: Vocoder framework used during the synthesis stage. Inputs are the PCA, F0 and MGC coefficients generated by the HMMs.

3.2 Subjective test results

Three voices were evaluated: Bruno (French male), kindly provided by Acapela Group, AWB (Scottish male) and SLT (US female) from the CMU ARCTIC database available in [12]. The training set had a duration of about 50 min for AWB and SLT, and 2 h for Bruno and was composed of phonetically balanced utterances sampled at 16 kHz. The subjective test was submitted to 20 non-professional listeners. It consisted of 4 synthesized sentences of about 7 seconds per speaker. For each sentence, two versions were presented (using either the traditional or proposed excitation) and the subjects were asked to vote for the one they preferred (if any). Averaged results are shown in Figure 9. A considerable gain through the use of our excitation is reached on the male voices. Although still present, this advantage turns out to be less dominant for the female speaker.

Figure 9: Results of the preference test.

4. CONCLUSIONS AND FUTURE WORKS

This paper proposed a new model of excitation for parametric speech synthesis. This model relies on the decomposition of pitch-synchronous excitation frames on a limited number (typically around 20-30) of eigenresiduals, obtained on a relatively small dataset (20 min are generally sufficient). This approach was tested on HMM-based synthesis by enhancing the source modeling with PCA-based coefficients. A substantial improvement was reported on male as well as female (even though less pronounced) speakers, while the system footprint still remains around 1Mb.

As future works, we plan to investigate the following directions:

- One of the main problems encountered in voice conversion could be alleviated by the proposed approach. Indeed, while such systems achieve their main goal, i.e the converted voice is recognized as being uttered by the target speaker, they suffer from a poor quality in the delivered speech [20]. Replacing the traditional pulse excitation by the use of eigenresiduals (beforehand computed for the target speaker) should also undoubtedly lead to significant improvements.
- Since the glottal source could differ with the phonetic context, analyzing the effect on a preliminary stage of clustering on the eigenresiduals would be worthwhile. More precisely, the HMM-based speech synthesizer makes use of binary decision trees achieving a Context-Oriented Clustering. We therefore plan to apply such a Principal Component Analysis for different nodes

of these trees, and observe whether a significant difference in the eigenresidual waveform is noticeable, and if this way of proceeding leads to an improvement in the quality of the synthesized speech (sufficiently important for the increase of complexity).

- Although we proposed the use of a Principal Component Analysis, other data mining methods (possibly derived from the functional PCA literature, [21]) could be efficiently employed to extract a suitable representation from the large dataset of normalized GCI-centered residual frames (obtained as described in Section 2.1).
- Finally, it would certainly be very interesting to compare the proposed approach with other techniques of excitation modeling, such as STRAIGHT [22], the mixed excitation [7],[8], or based on the Liljencrant-Fant model [9]. Although all these approaches reported a relative improvement with regard to the traditional pulse excitation, no comparison is available yet, since authors worked with different synthesis frameworks and with different databases.

5. ACKNOWLEDGMENTS

Thomas Drugman is supported by the "Fonds National de la Recherche Scientifique" (FNRS). Authors also would like to thank the reviewers for their helpful feedback.

REFERENCES

[1] A.J Hunt and A.W Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP96*, 1996, pp. 373-376.

[2] A.W. Black, H. Zen and K. Tokuda, "Statistical Parametric Speech Synthesis," in *Proc. ICASSP07*, 2007, pp. 1229-1232.

[3] K. Tokuda, H. Zen and A.W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Workshop on Speech Synthesis*, 2002, 227-230.

[4] H. Zen, T. Toda and K. Tokuda, "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006," in *IEICE Trans. on Information and Systems*, 2006.

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

[6] H. Zen, K. Tokuda and T. Kitamura, "An introduction of trajectory model into HMM-based speech synthesis," in *Proc. of 5th ISCA Speech Synthesis Workshop*, 2004.

[7] T. Yoshimura, K. Tokuda, T. Masuko and T. Kitamura, "Mixed-excitation for HMM-based speech synthesis," in *Proc. Eurospeech01*, 2001, pp. 2259-2262.

[8] R. Maia, T. Toda, H. Zen, Y.Nankaku and K. Tokuda, "An excitation model for HMM-based speech synthesis based on residual modeling," in *Proc. ISCA SSW6*, 2007.

[9] J. Cabral, S. Renals, K. Richmond and J. Yamagishi, "Glottal Spectral Separation for Parametric Speech Synthesis," in *Proc. Interspeech*, pp. 1829-1832, 2008.

[10] T. Drugman, G. Wilfart, A. Moinet and T. Dutoit, "Using a pitch-synchronous residual for hybrid HMM/frame selection speech synthesis," in *Proc. ICASSP09*, 2009.

[11] HMM-based Speech Synthesis System (HTS), [Online], *http://hts.sp.nitech.ac.jp/*

[12] CMU ARCTIC speech synthesis databases, [Online], *http://festvox.org/cmu_arctic/*

[13] K. Tokuda, T. Kobayashi, T. Masuko and S. Imai, "Mel generalized cepstral analysis A unified approach to speech spectral estimation," in *ICSLP*, 1994.

[14] H. Kawahara, Y. Atake and P. Zolfaghari, "Accurate vocal event detection method based on a fixedpoint analysis of mapping from time to weighted average group delay," in *ICSLP*, 2000, 664-667.

[15] J.P Cabral and L.C. Oliveira, "Pitch-Synchronous Time-Scaling for Prosodic and Voice Quality Transformations," in *Proc. Interspeech05*, 2005, pp. 1137-1140.

[16] I.T. Jolliffe, "Principal Component Analysis," in *Springer Series in Statistics*, 2nd ed., New-York, 2002

[17] K. Tokuda, T. Masuko, N. Myiazaki and T. Kobayashi, "Multi-space probability distribution HMM," in *IEICE Trans. on Information and Systems*, 2002, vol. E85-D, pp.455-464.

[18] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Duration modeling for HMM-based speech synthesis," in *Proc. ICSLP98*, 1998.

[19] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," in *Proc. Eurospeech95*, 1995.

[20] Y. Stylianou, "Voice transformation: a survey," in *Proc. ICASSP09*, pp.3585-3588, 2009.

[21] J. Ramsay and B. Silverman, "Functional data analysis," *Springer Series in Statistics*, 2nd edition, 2005.

[22] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino and H. Banno, "TANDEM-STRAIGHT: a Temporally Stable Power Spectral Representation for Periodic Signals and Applications to Interference-free Spectrum, F0, and Aperiodicity Estimation," in *Proc. ICASSP08*, pp.3933-3936, 2008.