A HMM-BASED SPEECH SYNTHESIS SYSTEM USING A NEW GLOTTAL SOURCE AND VOCAL-TRACT SEPARATION METHOD

Pierre Lanchantin, Gilles Degottex, Xavier Rodet

IRCAM - CNRS-UMR9912-STMS,
Analysis-Synthesis Team
1, place Igor-Stravinsky,
75004 Paris, France

ABSTRACT

This paper introduces a HMM-based speech synthesis system which uses a new method for the Separation of Vocal-tract and Liljencrants-Fant model plus Noise (SVLN). The glottal source is separated into two components: a deterministic glottal waveform Liljencrants-Fant model and a modulated Gaussian noise. This glottal source is first estimated and then used in the vocal-tract estimation procedure. Then, the parameters of the source and the vocal-tract are included into HMM contextual models of phonems. SVLN is promising for voice transformation in synthesis of expressive speech since it allows an independent control of vocal-tract and glottal-source properties. The synthesis results are finally discussed and subjectively evaluated.

Index Terms— HMM-based speech synthesis, Liljencrants-Fant model

1. INTRODUCTION

Emerging fields of speech applications such as avatars, spoken dialog system, cinema, or serious games require high quality and expressive speech. Unit-selection speech synthesis still provides the best available quality but speaker identity and expressivity are limited by the content of the underlying database even though slight transformations can be applied. Recently HMM-based speech synthesis has received great attention because it allows a precise control of speech attributes as well as the use of adaptation methods inherited from speech recognition allowing interpolation between speaker identity or expressivity.

Earlier HMM-based speech synthesis systems suffered from buzzy speech quality due to a simple excitation model involving a impulse train and white noise to model voiced and unvoiced segments respectively. Several methods have been proposed over recent years to improve the excitation model such as the use of the STRAIGHT vocoder [1] in order to shape a multi-band mixed excitation (ME) with the spectral envelope [2]. ME has also been used and improved in [3, 4, 5]. In these methods, the Vocal-Tract Filter (VTF) is usually assumed to be excited by a flat amplitude spectrum. The spectral amplitude of the source is thus merged into the VTF estimate. In [6], Cabral et al. proposed to use Liljencrants-Fant (LF) model and demonstrate the parametric flexibility of the LF-model for voice transformation. In [7], they proposed a global spectrum separation method to estimate the voice source and the VTF, estimating the latter by removing the spectral effects of the calculated glottal source model but without taking into account the different properties of the source (deterministic or stochastic).

In this paper, we present a HMM-based speech synthesis system for French using a new glottal source and vocal-tract separation method called Separation of Vocal-tract and Liljencrants-Fant model plus Noise (SVLN), different from the one described in [7]. The glottal excitation is separated into two additive components: a deterministic glottal waveform modeled by the LF model [8] and a stochastic component modeled by a Gaussian noise. The parametrization by only two parameters - the shape parameter Rd and the noise level σ_g - allows an intuitive control of the voice quality. An estimate of the LF model [9] is used to extract these parameters and the VTF is estimated by taking the estimate of the glottal source into account. The VTF parameters are thus independent of the excitation parameters and the glottal source may be changed, keeping the VTF untouched which can be of interest for voice transformations in expressive speech synthesis.

The paper is structured as follows: the SVLN method is described in Section 2. Synthesis is presented in Section 3. Those speech synthesis system components which are specific to French, and the integration of the new method into the system are described in Section 4. Finally, the system is subjectively evaluated and compared to the state-of-the-art systems in Section 5 and we conclude in Section 6.

2. GLOTTAL SOURCE AND VOCAL-TRACT SEPARATION AND ESTIMATION

The SVLN method is described in this Section. We first describe the acoustic waveform model, its separation and its parametrization and we give the estimation method for each of the parameters.

2.1. Speech model

The signal is assumed to be stationary in a short analysis window ($\approx$ 3 periods in voiced parts, $5ms$ in unvoiced parts). In the frequency domain, the voice production model of an observed speech spectrum $S(\omega)$ is (see Fig. 1):

$$S(\omega) = (H^{f_0}(\omega) \cdot G^{Rd}(\omega) + N^{\sigma_g}(\omega)) \cdot C^{\bar{c}}(\omega) \cdot L(\omega) \qquad (1)$$

H^{f_0} is a harmonic structure with fundamental frequency f_0. G^{Rd} is the deterministic excitation, i.e. an LF model [8]. This model is defined by: the fundamental period $1/f_0$, 3 shape parameters and the gain of the excitation E_e. To simplify the LF control, the parameter space is limited to a meaningful curve and a position defined by the value of a new parameter Rd [8, 10]. N^{σ_g} is a white Gaussian noise with standard deviation σ_g. C is the response of the VTF, a minimum-phase filter parametrized by cepstral coefficients $\bar{c}$ on a

mel scale. To avoid a dependency between the gains E_e and σ_g on one hand and the VTF mean amplitude on the other hand, a constraint is necessary. In this presentation, $G^{Rd}(\omega)$ is normalized by $G^{Rd}(0)$ and E_e is therefore unnecessary. Finally, L is the lips radiation. This filter is assumed to be the time derivative ($L(\omega) = j\omega$).

In the following Sections, $S(\omega)$ is the Fourier transform of a windowed speech signal according to the stationarity hypothesis given above. The model parameters $\{f_0, Rd, \sigma_g, \bar{c}\}$ are estimated on $S(\omega)$ at regular intervals of $5ms$ along the speech signal.

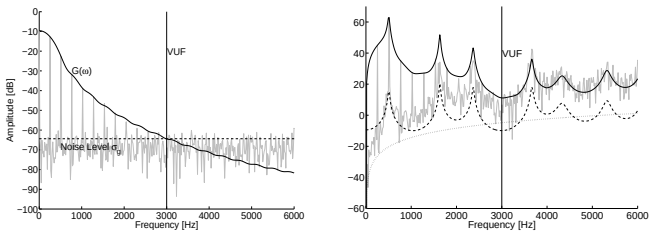

Fig. 1. The mixed excitation model on the left: G^{Rd} (solid line) and the noise level σ_g (dashed line). The speech model on the right: the VTF (dashed line); L (dotted line); the spectrum of one speech period (solid line). Both plots show in gray the source $(H \cdot G + N)$ or the speech spectrum $S(\omega)$ respectively.

2.2. Glottal source estimation

The fundamental frequency f_0 can be computed from the speech signal with a number of methods. For this presentation, we use a harmonic matching method [11]. To estimate the LF shape parameter Rd, the phase of the VTF is assumed to be close to zero in the range of the glottal formant frequencies. In this band, a minimum phase version of the LF model is fitted to a minimum phase envelope of $S(\omega)$ [9]. According to Fig. 1, an estimation of a Voiced/Unvoiced Frequency (VUF) is used to split $S(\omega)$ into a deterministic source below the VUF and white noise above [12]. Therefore, we assume that $G^{Rd}(\omega)$ crosses the noise mean amplitude at the VUF (Fig. 1). Consequently, knowing the spectral amplitude $|G^{Rd}(\omega)|$, σ_g can be deduced:

$$\sigma_g = |G^{Rd}(VUF)| \frac{\sqrt{2}}{\sqrt{\pi/2} \cdot \sqrt{\sum_t \mathrm{win}[t]^2}}$$

This is because $|G^{Rd}(\omega)|$ is the expected amplitude of the LF model and spectral amplitudes of Gaussian noise obey a Rayleigh distribution [11]. Consequently, $|G^{Rd}(VUF)|$ has to be converted to the Gaussian parameter σ_g through the Rayleigh mode ($\sqrt{2}/\sqrt{\pi/2}$) [11]. Additionally, in the spectral domain, the noise level is proportional to the energy of the analysis window used to compute $S(\omega)$, i.e. $\sqrt{\sum_t \mathrm{win}[t]^2}$.

2.3. Vocal-tract estimation

To estimate the VTF, the deterministic and stochastic frequency bands are modeled by two different envelopes according to their excitation properties. Then, these two envelopes are aligned with their expected amplitude. In the deterministic band, the contribution of the lips radiation $L(\omega)$ and the deterministic source $G^{Rd}(\omega)$ are removed from $S(\omega)$ by division in frequency (deconvolution in time) (eq. 2). Then, a cepstral envelope $\mathcal{T}^o$ of order o is fitted (by an iterative method [13]) on the top of the harmonic partials of the division result. $\mathcal{T}^o$ fits the expected amplitude of the excitation since the top of a harmonic partial is the expected amplitude. In the stochastic band, $S(\omega)$ is divided by $L(\omega)$ and by the crossing value $G^{Rd}(VUF)$ to assure a continuity between the two bands. Then, the division result is modeled by computing its power cepstrum $\mathcal{C}^o$ truncated to a given order o. According to the Rayleigh distribution, the expected amplitude of this frequency band is retrieved through the mean log amplitude measured by $\mathcal{C}^o$ ($\sqrt{\pi/2}/e^{0.058}$ in eq. 2) [11].

$$C(\omega) = \begin{cases} \mathcal{T}^o\left(\frac{S(\omega)}{L(\omega)G^{Rd}(\omega)}\right) \cdot \gamma^{-1} & \text{if } \omega < VUF \\ \mathcal{C}^o\left(\frac{S(\omega)}{L(\omega)G^{Rd}(VUF)}\right) \cdot \frac{\sqrt{\pi/2}}{\gamma \cdot e^{0.058}} & \text{if } \omega \geq VUF \end{cases} \quad (2)$$

In the synthesis step, the VTF is applied on each period. Therefore, its gain is normalized by the quantity of periods in the analysis window $\gamma = \sum_t \mathrm{win}[t]/(f_s/f_0)$ (f_s is the sampling frequency). The order of the envelopes $\mathcal{T}^o$ and $\mathcal{C}^o$ is chosen to avoid the fitting of the harmonic structure H^{f_0}. Therefore, $o = 0.5 \cdot f_s/f_0$. Even if no harmonic partial appears in the stochastic part, partials with distance of f_0 (but not multiples of f_0) appear in this part because the glottal noise is amplitude modulated by the glottal area [10]. Finally, to avoid the division by zero by $L(\omega)$ at zero frequency, $L(\omega)$ is replaced by $1 - \mu e^{j\omega}$ with μ close to unity. The mel cepstral coefficients $\bar{c}$ are computed from $C(\omega)$.

In unvoiced segments, no glottal pulses are synthesized. When VUF is lower than f_0, VUF is clipped to zero and f_0 is fixed to zero indicating an unvoiced segment.

3. SYNTHESIS BY SEGMENTS

The synthesis process is an overlap-add method: First, small segments of stationary signals are generated. Then, these segments are overlap-added to construct the whole signal.

In voiced parts ($f_0 > 0$), temporal marks m_k (Fig. 2) are synthesized with intervals corresponding to the fundamental period $1/f_0$. The maximum excitation instant [8] of the LF model is placed on this mark. Then, we define the starting time t_k of the k^{th}-segment as the opening instant [8] of the LF model. Finally, the ending time of the k^{th}-segment is the starting time of the next segment. In unvoiced parts ($f_0 = 0$), we use segments of $5ms$ and the mark m_k is placed in the middle (Fig. 2).

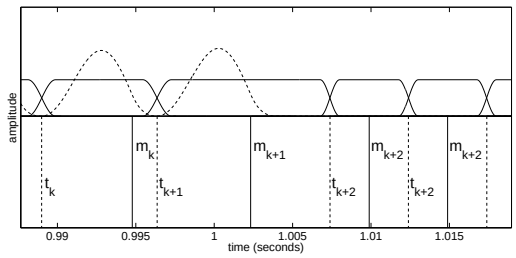

Fig. 2. 2 voiced segments followed by 2 unvoiced segments: synthesized LF models (dashed lines) and Windows win_k (solid lines). Marks and starting times in vertical lines.

For the deterministic excitation, no window is necessary to cross fade the glottal pulses since they start and end at zero time amplitude. Conversely, the noise is a continuous excitation. To control the noise amplitude (with σ_g) as its color (see below), a cross-fade has to be

used between segments. For each k^{th}-segment, a window win_k is thus built with a fade-in center on t_k and a fade-out center on t_{k+1} (Fig. 2). The fade-in/out function is a half Hanning window of duration $0.25 \cdot \min(t_{k+1} - t_k, t_k - t_{k-1})$. Additionally, the fade-out of win_k is the complementary of the fade-in of win_{k+1}. Consequently, the sum of all windows is 1 at any time. To improve the naturalness of the glottal noise in voiced segments, two processes are applied. First, the noise is high-pass filtered with a cutoff frequency equal to the VUF (in synthesis, the VUF is retrieved from the crossing value between $G^{Rd_k}(\omega)$ and σ_g). With such a filtering the lowest harmonics are not disturbed by the noise. Secondly, the noise is amplitude modulated with a function $v^{Rd}[t]$ built from the LF model as proposed in [14].

In voiced segments, the source of the k^{th}-segment is:

$$\begin{aligned} E_k(\omega) &= e^{j\omega m_k} \cdot G^{Rd_k}(\omega) \\ &\quad + F_{hp}^{VUF_k}(\omega) \cdot \mathcal{F}\big(v^{Rd_k}[t] \cdot \mathrm{win}_k[t] \cdot n^{\sigma_{gk}}[t]\big) \end{aligned}$$

where $n^{\sigma_k}[t]$ is a Gaussian random time sequence, $\mathcal{F}(.)$ is the Discrete Time Fourier Transform and F_{hp}^{VUF} is the high-pass filter of the noise. In unvoiced segments, the source reduces to:

$$E_k(\omega) = \mathcal{F}\big(\mathrm{win}_k[t] \cdot n^{\sigma_{gk}}[t]\big)$$

Finally, the speech spectrum is $S_k(\omega) = E_k(\omega) \cdot C^{\bar{c}_k}(\omega) \cdot j\omega$ and the whole signal is synthesized by overlap-adding the speech segments in the time domain.

4. INTEGRATION IN THE HMM-BASED SPEECH SYNHESIS SYSTEM

The implementation of our HMM-based speech synthesis system is based on the HTS Toolkit[15]. Parameters $\{f_0, Rd, \sigma_g, \bar{c}\}$ with $card(\bar{c}) = 32$ are estimated according to the method described in Section 2. Several stream configurations were tested and we finally kept the following one:

- one single Gaussian distribution with semi-tied covariance [16] for $\{Rd, \sigma_g, \bar{c}\}$;
- one multi-space distribution (MSD[17]) for f_0

Both streams include first and second derivatives of their parameters. In the second stream, parameters are modeled by a single Gaussian distribution with diagonal covariance for voiced parts, and the voiced/unvoiced decision is taken into account by a specific weight applied on each space in the MSD.

The naturalness of a synthetic voice also depends on the choice of the context features. We used the context features describing the phonetic context and lexical and syntactic features predicted from the text, as detailed in Table 1. These features have been automatically extracted from speech recordings and their text transcriptions using ircamAlign [18], an HMM-based segmentation system relying on HTK toolkit [19] and the Lia_phon [20] French phonetizer. The French text is first converted into a phonetic graph allowing multiple pronunciations. Then, the best phonetic sequence is chosen according to the audio file and aligned temporally on it. The context features are extracted according to the text and the extracted phonetic sequence. Finally, a 5-states left-to-right HMM was used to model each contextual phoneme.

The training procedure is similar to the one described in [21] : monophones models are first trained and then converted to context dependent models. Decision tree clustering is performed according to the extracted context features in order to robustly estimate the model parameters. During the synthesis step, parameters are first generated by HTS using a constrained maximum likelihood algorithm [22] from which a speech signal is synthesized according to the method described in section 3.

Phonetic features:

- **Phoneme identity**, and some phonological features (vowel lenght/height/fronting/rounding consonant type/place/voicing) in quintphone context

Lexical and syntactic features

- **Phoneme and syllabe structure**: pos-in-syl, syl-numsegs, syl-numsegs-{prev,next}, pos-in-word, pos-in-phrase, syl-nucleus
- **Word related**: word-POS-{prev,curr,next}, word-numsyl{prev,curr,next}, contentwords-from-phrase-{start,end}, words-from-contentword{prev,next}
- **Phrase related**: phrase-numsyls, phrase-numwords, pos-in-utterance
- **Utterance related** : Utt-numsyls, Utt-numwords, Utt-numphrases
- **Punctuation related**: phrase-punct

Table 1. List of the context features extracted by ircamAlign

5. EXPERIMENTS

The proposed system has been trained on a database of 1995 sentences (approximately 1h30 of speech) spoken by a French non professional male speaker and recorded at 16kHz in an anechoic room. The context features were automatically extracted from the database using ircamAlign as described in the previous Section. Three different systems are compared: simple pulse train excitation model, STRAIGHT and SVLN based ones. The two first ones were chosen for the comparison because their quality is generally well known and STRAIGHT is generally considered as a reference one. The three systems were trained on the same database and the same extracted context features.

5.1. Subjective evaluation

The test consists of a subjective comparison between the 3 systems. A comparison category rating (CCR[23]) test was used to assess the quality of the synthetic speech generated by the SVLN-based system in comparison to synthetic speech generated by the pulse train and STRAIGHT-based systems. 5 sentences were chosen to generate the test samples for each system. 48 French naive listeners compared a total of 15 speech sample pairs. They were asked to attribute a score to the quality of the second sample of a pair compared to the quality of the first one on the comparison mean opinion score (CMOS) scale.

The test is available on [24]. The prosody suffers by defaults partly due to the nonnatural diction (over-articulated) of the non professional speaker. The ranking of the three systems was evaluated by averaging the scores of the CCR test for each method (shown on Figure 3). The results show that the system based on the proposed SVLN method provides a better quality than the one based on the pulse train method but not as good as the one based on STRAIGHT. We suppose that the difference of quality may be partly explained by some artefacts due to the sensitivity of the voicing detection (through the VUF estimate) and the noise filtering and modulation. This potential way of improvement to our system will be studied in the future.

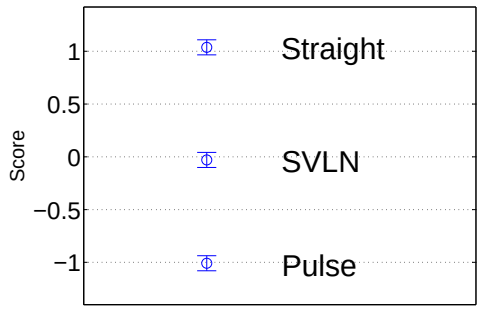

Fig. 3. Ranking of the CCR test for the following speech samples: STRAIGHT-based system, SVLN-based system, Pulse train excitation-based system. The 95% confidence intervals are presented for each score.

5.2. Voice Transformations

Although the quality of the speech synthesized by our system is not as good as the one synthesized by the STRAIGHT-based one, our system offers a better control of the vocal quality with only a few parameters. Due to the VTF parameters being independent of the excitation ones, the glottal source may be changed keeping the VTF untouched. For instance, in pitch transposition, the glottal formant [8] may be shifted independently of the VTF formants. This is of great value for expressive speech synthesis. Also it allows one to quickly synthesize different speaker personalities with various voice qualities from the same voice. Some voice tranformations examples are available on [24].

6. CONCLUSION

In this work, we proposed a HMM-based speech synthesis system for French using a new Separation method of Vocal-tract and Liljencrants-Fant model plus Noise (SVLN). With this method the naturalness obtained by HMM-based synthesis using a phase model like STRAIGHT can be approached with a better control of the vocal quality with few parameters. It also allows an intuitive control of the voice quality which is of great interest for expressive speech synthesis or to quickly synthesize different speaker personalities with various voice qualities from the same voice.

Future research includes the improvement of the analysis stability and robustness (e.g. against high frequency artefacts). We will also improve the prosody modelling by extracting more advanced context features using a French lexical analyser. Finally, we plan to use the SVLN method for voice conversion.

7. ACKNOWLEDGMENTS

This project was supported by the ANR project *Affective Avatars* and by a CNRS grant.

8. REFERENCES

[1] H. Kawahara, I Masuda-Katsuse, and A. Cheveigné, "Restructurin speech representations using a pitch-adaptative time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," in *Speech Communication*, 1999, vol. 27.

[2] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker independent HMM-based speech synthesis system - hts-2007 system for the blizzard challenge 2007," in *Proc of the Blizzard Challenge 2007*, 2007.

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

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

[5] T. Drugman, G. Wilfart, and T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *Interspeech09*, Brighton, U.K, 2009.

[6] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis," in *ISCA SSW6*, 2007.

[7] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Glottal spectral separation for parametric speech synthesis," in *Proc. Interspeech08*, 2008.

[8] G. Fant, "The LF-model revisited. transformations and frequency domain analysis.," *STL-QPSR*, vol. 36, no. 2-3, pp. 119–156, 1995.

[9] G. Degottex, A. Roebel, and X. Rodet, "Shape parameter estimate for a glottal model without time position," in *SPECOM*, 2009, pp. 345–349.

[10] H.-L. Lu, *Toward a High-quality Singing Synthesizer with Vocal Texture Control*, Ph.D. thesis, Stanford, 2002.

[11] C. Yeh, *Multiple fundamental frequency estimation of polyphonic recordings*, Ph.D. thesis, UPMC - Paris VI, juin 2008.

[12] S.-J. Kim and M. Hahn, "Two-band excitation for HMM-based speech synthesis," *IEICE - Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 378–381, 2007.

[13] A. Roebel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *DAFx*, 2005.

[14] A. del Pozo and S. Young, "The linear transformation of lf glottal waveforms for voice conversion," in *Interspeech*, 2008.

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

[16] M.J.F. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.

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

[18] P. Lanchantin, A.C. Morris, X. Rodet, and C. Veaux, "Automatic phoneme segmentation with relaxed textual constraints," in *LREC'08 Proceedings*, Marrakech, Morocco, 2008.

[19] S.J. Young and Sj Young, "The HTK hidden markov model toolkit: Design and philosophy," *Entropic Cambridge Research Laboratory, Ltd*, vol. 2, pp. 2–44, 1994.

[20] F. Bechet, "Liaphon : un système complet de phonetisation de textes," *Traitement Automatique des Langues*, vol. 42, no. 1, 2001.

[21] 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.

[22] 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 Eurospeech'95*, 1995.

[23] Recommendation ITU-U P.800, "Methods for subjective determination of transmission quality," in *International Telecommunication Union*, Aug. 1996.

[24] P. Lanchantin, G. Degottex, and X. Rodet, "Demonstration page of a HMM-based speech synthesis system using a new glottal source and vocal-tract separation method," `http://recherche.ircam.fr/anasyn/lanchant`.