Harmonic Model for Female Voice Emotional Synthesis

Anna Přibilová[1] and Jiří Přibil[2,3]

[1] Department of Radio Electronics, Slovak University of Technology
Ikovičova 3, SK-812 19 Bratislava, Slovakia
Anna.Pribilova@stuba.sk
[2] Institute of Photonics and Electronics, Academy of Sciences of the Czech Republic
Chaberská 57, CZ-182 51 Prague 8, Czech Republic
[3] Institute of Measurement Science, Slovak Academy of Sciences
Dúbravská cesta 9, SK-841 04 Bratislava, Slovakia
Jiri.Pribil@savba.sk

Abstract. Spectral and prosodic modifications for emotional speech synthesis using harmonic modelling are described. Autoregressive parameterization of inverse Fourier transformed log spectral envelope is used. Spectral flatness determines the voicing transition frequency dividing spectrum of synthesized speech into minimum phases and random phases of the harmonic model. Female emotional voice conversion is evaluated by a listening test.

Keywords: emotional speech, spectral envelope, harmonic speech model, emotional voice conversion.

1 Introduction

Expression of emotional states in human voice has been moved to the centre of attention of researchers involved in speech processing [1-8]. Our contribution to this area consists of female emotional voice conversion using harmonic sine-wave model of speech signal, i.e. a sinusoidal model with harmonically related sine waves [9, 10]. Although this model had originally been used in speech coding, its modifications were also successfully applied in speech synthesis [11-13]. For modelling of voiced fricatives and other speech sounds with mixed excitation the sine-wave phases are made random above the voicing transition frequency, which is determined by the voicing probability that is a measure of how well the harmonic set of sine waves fits the measured set of sine waves minimizing the mean squared error [9, 10]. In [13] the notion of a maximum voiced frequency is used for the same variable, however, the upper band of the spectrum is modelled using an all-pole filter driven by a white Gaussian noise instead of a sum of sine waves with random phases. A rather elaborate technique for decomposition of voiced speech into periodic and aperiodic components is described in [14].

Sinusoidal model has also been used for emotional speech analysis [15, 16]. Our approach to harmonic speech modelling with autoregressive (AR) parameterization of spectral envelope is described in Section 2. Modification of AR parameters according to different emotions is described in Section 3. Prosody modification is dealt with in Section 4. Listening test results are summarized in Section 5.

J. Fierrez et al. (Eds.): BioID_MultiComm2009, LNCS 5707, pp. 49–56, 2009.

2 Harmonic Speech Model with AR parameterization

Speech signal synthesized by the harmonic speech model with AR parameterization (Fig. 1) is represented by a sum of sine waves with frequencies $\{f_m\}$ corresponding to pitch harmonics, amplitudes $\{A_m\}$ given by sampling of the spectral envelope at these frequencies, and phases $\{\varphi_m\}$ being samples of Hilbert transform of the log spectral envelope corresponding to the minimum-phase model. For unvoiced speech and for voiced speech above the voicing transition frequency ($f_{v\text{-}uv}$) the phases are randomized in the interval $<-\pi, \pi>$. Our approach to $f_{v\text{-}uv}$ determination uses spectral flatness S_F [17] as a measure of degree of voicing. The spectral envelope is represented by AR parameters (gain G and LPC coefficients $\{a_n\}$). Summed sine waves in two consecutive pitch periods are weighted by an asymmetric window in such a way that the left part of the current asymmetric window has the same length as the right part of the previous window, and the right part of the current window has the same length as the left part of the next window, and the overlapped asymmetric windows are complementary. For the final synthesis the weighted overlapped consecutive pairs of pitch-synchronous frames are added to avoid discontinuities at the frame boundaries.

Speech analysis is performed in equidistant overlapping weighted frames according to the block diagram in Fig. 2. To avoid disadvantages of standard AR modelling (bias of formant frequencies toward pitch harmonics, underestimation of formant

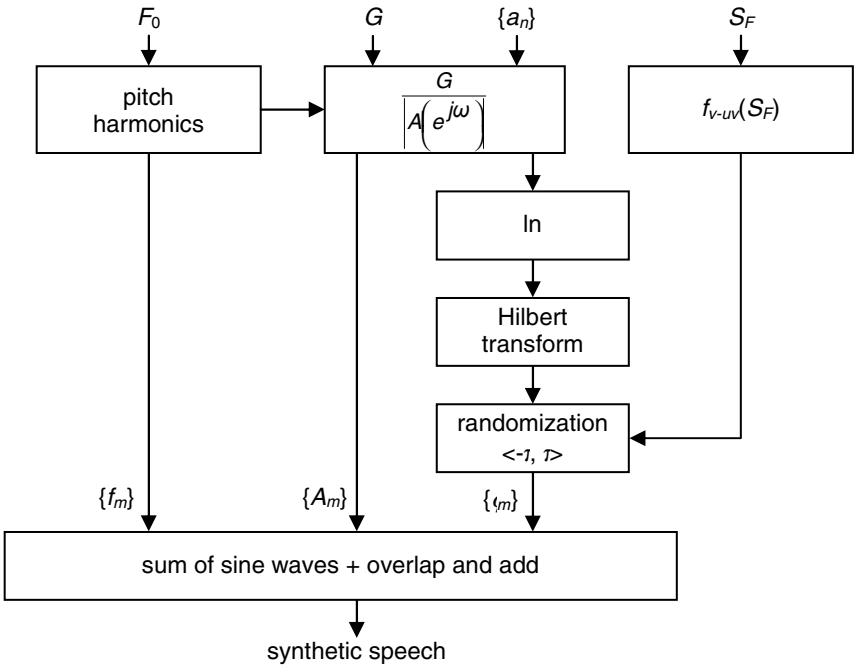

Fig. 1. Block diagram of the harmonic speech model with AR parameterization

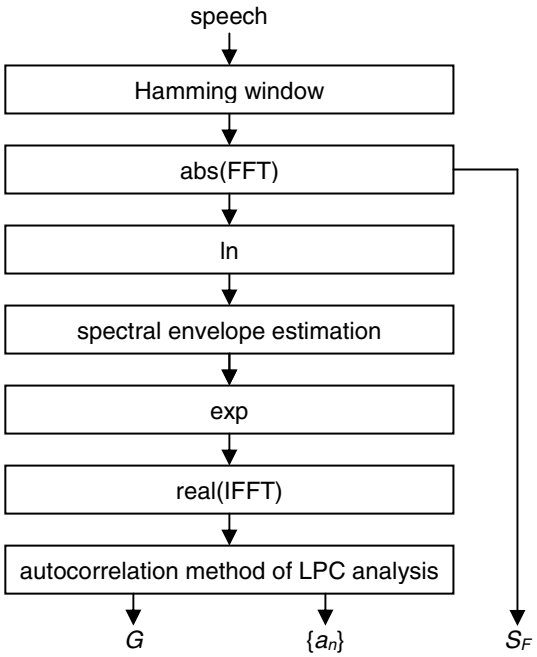

Fig. 2. Determination of model parameters in one equidistant speech frame

bandwidth) the AR parameters are computed from the time-domain signal corresponding to the spectral envelope instead of the original speech signal. Spectral envelope estimation similar to that of [18] is used. We use spline interpolation [19] applied to local maxima at pitch harmonics of the log spectrum.

3 Spectral Modifications for Emotional Synthesis

According to [20] larynx and pharynx expansion, vocal tract walls relaxation, and mouth corners retraction upward lead to falling first formant and rising higher formants during pleasant emotions. On the other hand, larynx and pharynx constriction, vocal tract walls tension, and mouth corners retraction downward lead to rising first formant and falling higher formants for unpleasant emotions. Thus, the first formant and the higher formants of emotional speech shift in opposite directions in the frequency ranges divided by a frequency between the first and the second formant.

Although the formant frequencies differ to some extent for different languages and their ranges are overlapped [21] the male voice vowel formant areas without overlap can be determined: $F_1 \approx 250 \div 700$ Hz, $F_2 \approx 700 \div 2000$ Hz, $F_3 \approx 2000 \div 3200$ Hz, $F_4 \approx 3200 \div 4000$ Hz [22]. Using the general knowledge of [21] that females have on average 20 % higher formant frequencies than males, female voice vowel formant areas without overlap will be: $F_1 \approx 300 \div 840$ Hz, $F_2 \approx 840 \div 2400$ Hz, $F_3 \approx 2400 \div 3840$ Hz, $F_4 \approx 3840 \div 4800$ Hz. The border frequency between the first and the second formant for female voice will be $F_{1,2} = 840$ Hz.

Spectral parameters modification consists of spectral envelope modification by non-linear frequency scale transformation of the spectral envelope computed using AR parameters obtained during analysis. After spectral transformation, the inverse Fourier transform of the spectral envelope is treated as a real speech signal for modified AR parameters computation for a database of AR parameters corresponding to different emotions – see Fig. 3.

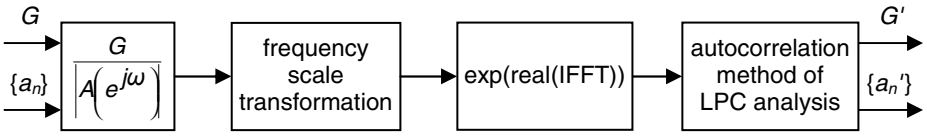

Fig. 3. Modification of AR parameters using frequency scale transformation

For shifting of the first formant and the higher formants in the opposite directions we use a smooth function of frequency representing formant ratio between emotional and neutral speech. For its better analytic representation the frequency scale is logarithmically warped so that the border frequency $F_{1,2}$ corresponds to one fourth of the sampling frequency f_s. Inverse of this log warping function is

$$f(f_t) = a\, b^{f_t} + c\,, \tag{1}$$

where f represents the input frequency and f_t corresponds to the transformed frequency. Unknown variables a, b, c are determined using the points $[f_t, f] = [0, 0]$, $[f_s/4, F_{1,2}]$, $[f_s/2, f_s/2]$. The solution of the system of the three equations is

$$a = \frac{2F_{1,2}^2}{f_s - 4F_{1,2}}, \qquad b = \exp\left(\frac{4\ln\left(\dfrac{f_s - 2F_{1,2}}{2F_{1,2}}\right)}{f_s}\right), \qquad c = -\,a. \tag{2}$$

Inversion of (1) gives the logarithmically warped frequency scale

$$f_t(f) = \log_b \frac{f-c}{a} = \frac{\ln\left(\dfrac{f-c}{a}\right)}{\ln b}. \tag{3}$$

Formant ratio $\gamma(f_t)$ as a smooth function of the logarithmically warped frequency can be expressed by a fourth-order polynomial function

$$\gamma(f_t) = p\, f_t^{\,4} + q\, f_t^{\,3} + r\, f_t^{\,2} + s\, f_t + t. \tag{4}$$

Coefficients of this polynomial are computed from equidistant points $[f_t, \gamma] = [0, 1]$, $[f_s/8, \gamma_1]$, $[f_s/4, 1]$, $[3f_s/8, \gamma_2]$, $[f_s/2, 1]$. Solution of the system of five equations is

$$p = \frac{2048}{3f_s^4}(-\gamma_1 - \gamma_2 + 2), \qquad q = \frac{256}{3f_s^3}(9\gamma_1 + 7\gamma_2 - 16),$$
$$r = \frac{64}{3f_s^2}(-13\gamma_1 - 7\gamma_2 + 20), \qquad s = \frac{32}{3f_s}(3\gamma_1 + \gamma_2 - 4), \qquad t = 1. \tag{5}$$

Relation between the modified spectral envelope $E'(f)$ and the original one $E(f)$ is

$$E'(f) = E\left(\frac{f}{\gamma(f_t(f))}\right). \tag{6}$$

For 16-kHz sampling the transformation function (3) gives the frequency $f_s/8$ corresponding to 214.3 Hz and the frequency $3f_s/8$ corresponding to 2666.7 Hz. Chosen female emotional-to-neutral formant ratios at these frequencies together with spectral flatness ratio obtained by emotional speech analysis are shown in Table 1.

Table 1. Emotional-to-neutral formant ratios γ_1, γ_2 and spectral flatness ratio S_F

	γ_1	γ_2	S_F
joyous-to-neutral	0.70	1.05	1.24
angry-to-neutral	1.35	0.85	1.11
sad-to-neutral	1.10	0.90	2.02

In the four vowel formant areas the mean formant ratios are computed using the formant transformation function (4). Their values are shown in Table 2. For joy the first formant is shifted to the left by about 10 %, the second and third formants are shifted to the right by about 3 % to 6 % and the shift gradually decreases. For anger the first formant is shifted to the right by about 13 %, the higher formants are shifted to the left by about 10 % to 14 %. For sadness the mean shift of the first formant is about 4 % to the right and the higher formants about 6 % to 10 % to the left.

Table 2. Mean female emotional-to-neutral formant ratios in formant areas for chosen γ_1, γ_2

	300÷840 Hz	840÷2400 Hz	2400÷3840 Hz	3840÷4800 Hz
joyous-to-neutral	0.8982	1.0589	1.0334	0.9964
angry-to-neutral	1.1289	0.8849	0.8623	0.9012
sad-to-neutral	1.0432	0.9383	0.8991	0.9076

4 Prosodic Modifications for Emotional Synthesis

For emotional speech conversion, following prosodic parameters are modified: F0 mean, F0 range, energy, and duration. For joyous/angry emotional styles rising/falling *linear trend* (LT) of F0 is used at the end of sentences. Modification ratios between emotional and neutral speech were chosen experimentally as shown in Table 3.

Table 3. Prosodic parameters modification ratio values between emotional and neutral speech

	F0 mean	F0 range	energy	duration	LT type	LT_{start}
joy	1.18	1.30	1.30	0.81	rising	55 %
anger	1.16	1.30	1.70	0.84	falling	35 %
sadness	0.81	0.62	0.95	1.16	–	0

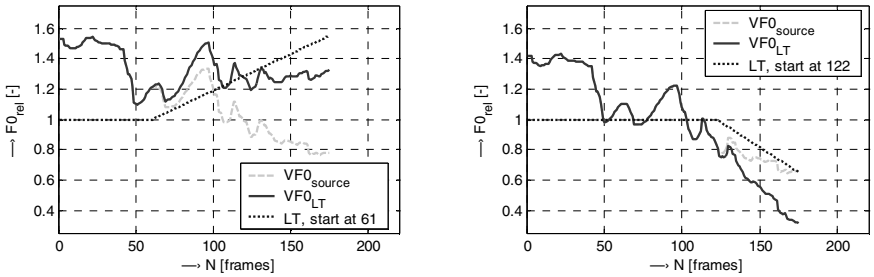

Fig. 4. Linear trend applied to VF0 contour for joyous (left) and angry (right) emotional styles. Source VF0 contour normalized by $F0_{mean}$ in the sentence "Vše co potřeboval" ("All he needed") – female speaker, f_s = 16 kHz, frame length = 8 ms.

Applying of LT to *virtual F0* (VF0) contour obtained by cubic interpolation in unvoiced parts of speech can be seen in Fig. 4. Starting point of LT is determined by a parameter LT_{start} (in percentage of distance to the end of the sentence).

5 Listening Tests

Subjective evaluation called "Determination of emotion type" was realized by the listening test with the help of automated listening test program located on the web page http://www.lef.um.savba.sk/scripts/itstposl2.dll. Every listening test consists of ten evaluation sets selected randomly from the testing corpus composed of 60 short sentences with durations varying between 1 and 3.5 seconds. The sentences were extracted from the Czech stories narrated by a female professional actor. For each sentence there is a choice from four possibilities: "joy", "sadness", "anger", or "other".

Twenty listeners (16 Czechs and 4 Slovaks, 6 women and 14 men) took part in the listening test. The summary results are presented in the form of a confusion matrix in Table 4. Best identified is sadness, worst identified is joy. Evaluation of successful determination of emotion type in individual sentences was carried out, too. Table 5 shows summed relative values for all emotions (values in the column "not classified" represent choice "other" in the listening test, "exchanged" corresponds to incorrect choice).

Table 4. Confusion matrix of the listening test

	joy	anger	sadness	other
joy	**59.0 %**	0.5 %	16.0 %	24.5 %
anger	2.5 %	**73.5 %**	2.0 %	22.0 %
sadness	0.5 %	0.5 %	**90.0 %**	9.0 %

Table 5. Best and worst evaluated sentences (data summed for all emotions)

	sentence	correct	not classified	exchanged
best evaluated	s13*	88.1 %	11.9 %	0 %
worst evaluated	s12**	57.6 %	30.3 %	12.1 %

* "Vše co potřeboval." ("All he needed.")

** "Máš ho mít." ("You ought to have it.")

6 Conclusion

Performed listening tests have shown that combination of spectral and prosodic modification in female voice emotion conversion using harmonic speech modelling gives the best results for sadness and the worst results for joy. The best listening test results correspond to the sentences that had been uttered most neutrally in the original. Some sentences of the original speech extracted from the stories were emotionally coloured, and then, their resyntheses with emotional modifications were perceptually biased towards the original emotion.

In our next research, we want to include microprosodic features in emotional voice conversion and we intend to experiment with application of linear trend F0 modifications also at the beginning of sentences.

Acknowledgment. This work has been done in the framework of the COST Action 2102. It has also been supported by the Ministry of Education of the Slovak Republic (MVTS COST 2102/STU/08), the Ministry of Education, Youth, and Sports of the Czech Republic (OC08010), the Grant Agency of the Czech Republic (GA102/09/0989), and the Grant Agency of the Slovak Academy of Sciences (VEGA 2/0142/08).

References

1. Navas, E., Hernáez, I., Luengo, I.: An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS. IEEE Transactions on Audio, Speech, and Language Processing 14, 1117–1127 (2006)
2. Tao, J., Kang, Y., Li, A.: Prosody Conversion from Neutral Speech to Emotional Speech. IEEE Transactions on Audio, Speech, and Language Processing 14, 1145–1154 (2006)
3. Ververidis, D., Kotropoulos, C.: Emotional Speech Recognition: Resources, Features, and Methods. Speech Communication 48, 1162–1181 (2006)

4. Tóth, S.L., Sztahó, D., Vicsi, K.: Speech Emotion Perception by Human and Machine. In: Esposito, A., Bourbakis, N.G., Avouris, N., Hatzilygeroudis, I. (eds.) HH and HM Interaction. LNCS (LNAI), vol. 5042, pp. 213–224. Springer, Heidelberg (2008)
5. Zainkó, C., Fék, M., Németh, G.: Expressive Speech Synthesis Using Emotion-Specific Speech Inventories. In: Esposito, A., Bourbakis, N.G., Avouris, N., Hatzilygeroudis, I. (eds.) HH and HM Interaction. LNCS (LNAI), vol. 5042, pp. 225–234. Springer, Heidelberg (2008)
6. Kostoulas, T., Ganchev, T., Fakotakis, N.: Study on Speaker-Independent Emotion Recognition from Speech on Real-World Data. In: Esposito, A., Bourbakis, N.G., Avouris, N., Hatzilygeroudis, I. (eds.) HH and HM Interaction. LNCS (LNAI), vol. 5042, pp. 235–242. Springer, Heidelberg (2008)
7. Ringeval, F., Chetouani, M.: Exploiting a Vowel Based Approach for Acted Emotion Recognition. In: Esposito, A., Bourbakis, N.G., Avouris, N., Hatzilygeroudis, I. (eds.) HH and HM Interaction. LNCS (LNAI), vol. 5042, pp. 243–254. Springer, Heidelberg (2008)
8. Callejas, Z., López-Cózar, R.: Influence of Contextual Information in Emotion Annotation for Spoken Dialogue Systems. Speech Communication 50, 416–433 (2008)
9. McAulay, R.J., Quatieri, T.F.: Low-Rate Speech Coding Based on the Sinusoidal Model. In: Furui, S., Sondhi, M.M. (eds.) Advances in Speech Signal Processing, pp. 165–208. Marcel Dekker, New York (1992)
10. McAulay, R.J., Quatieri, T.F.: Sinusoidal Coding. In: Kleijn, W.B., Paliwal, K.K. (eds.) Speech Coding and Synthesis, pp. 121–173. Elsevier Science, Amsterdam (1995)
11. Dutoit, T., Gosselin, B.: On the Use of a Hybrid Harmonic/Stochastic Model for TTS Synthesis-by-Concatenation. Speech Communication 19, 119–143 (1996)
12. Bailly, G.: Accurate Estimation of Sinusoidal Parameters in a Harmonic+Noise Model for Speech Synthesis. In: Eurospeech 1999, Budapest, pp. 1051–1054 (1999)
13. Stylianou, Y.: Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis. IEEE Transactions on Speech and Audio Processing 9, 21–29 (2001)
14. Yegnanarayana, B., d'Alessandro, C., Darsinos, V.: An Iterative Algorithm for Decomposition of Speech Signals into Periodic and Aperiodic Components. IEEE Transactions on Speech and Audio Processing 6, 1–11 (1998)
15. Drioli, C., Tisato, G., Cosi, P., Tesser, F.: Emotions and Voice Quality: Experiments with Sinusoidal Modeling. In: Proceedings of Voice Quality, Geneva, pp. 127–132 (2003)
16. Ramamohan, S., Dandapat, S.: Sinusoidal Model-Based Analysis and Classification of Stressed Speech. IEEE Transactions on Audio, Speech, and Language Processing 14, 737–746 (2006)
17. Gray, A.H., Markel, J.D.: A Spectral-Flatness Measure for Studying the Autocorrelation Method of Linear Prediction of Speech Analysis. IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP 22, 207–217 (1974)
18. Vích, R., Vondra, M.: Speech Spectrum Envelope Modeling. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) COST Action 2102. LNCS (LNAI), vol. 4775, pp. 129–137. Springer, Heidelberg (2007)
19. Unser, M.: Splines. A Perfect Fit for Signal and Image Processing. IEEE Signal Processing Magazine 16, 22–38 (1999)
20. Scherer, K.R.: Vocal Communication of Emotion: A Review of Research Paradigms. Speech Communication 40, 227–256 (2003)
21. Fant, G.: Acoustical Analysis of Speech. In: Crocker, M.J. (ed.) Encyclopedia of Acoustics, pp. 1589–1598. John Wiley & Sons, Chichester (1997)
22. Fant, G.: Speech Acoustics and Phonetics. Kluwer Academic Publishers, Dordrecht (2004)