Spectrum Modification for Emotional Speech Synthesis

Anna Přibilová[1] and Jiří Přibil[2]

[1] Department of Radio Electronics, Slovak University of Technology
Ilkovičova 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, Czech Republic
Jiri.Pribil@savba.sk

Abstract. Emotional state of a speaker is accompanied by physiological changes affecting respiration, phonation, and articulation. These changes are manifested mainly in prosodic patterns of F0, energy, and duration, but also in segmental parameters of speech spectrum. Therefore, our new emotional speech synthesis method is supplemented with spectrum modification. It comprises non-linear frequency scale transformation of speech spectral envelope, filtering for emphasizing low or high frequency range, and controlling of spectral noise by spectral flatness measure according to knowledge of psychological and phonetic research. The proposed spectral modification is combined with linear modification of F0 mean, F0 range, energy, and duration. Speech resynthesis with applied modification that should represent joy, anger and sadness is evaluated by a listening test.

Keywords: emotional speech, spectral envelope, speech synthesis, emotional voice conversion.

1 Introduction

Vocal emotion communication has been investigated by psychologists, but also psychiatrists, for a long time. Emotional arousal of a speaker is accompanied by physiological changes that affect respiration, phonation, and articulation and produce emotion-specific patterns of acoustic parameters. These acoustic changes are transmitted to the ears of the listener and perceived via the auditory perceptual system [1]. Information and communication technology development extended emotion recognition from humans to computers [2], [3], [4], [5], and emotion expression is also used to improve naturalness of synthetic speech [6].

Influence of emotions on human speech is manifested mainly in prosody [7], however, emotional state of a speaker is accompanied also by physiological changes causing shift of individual formants, different amount of low-frequency and high-frequency energy, and/or different spectral noise depending on the expressed emotion [1]. Our present work is aimed mainly at modification of these spectral parameters using known results of psychological and phonetic research. Prosody conversion is done by modification of neutral prosodic parameters according to known ratios between emotional and neutral speech.

A. Esposito et al. (Eds.): Multimodal Signals, LNAI 5398, pp. 232–241, 2009.
© Springer-Verlag Berlin Heidelberg 2009

2 Spectral Parameters Modification

Our approach to spectral parameters modification consists of spectral envelope modification by non-linear frequency scale transformation, spectral energy distribution modification by shelving filtering, and spectral noise modification by controlling the degree of voicing in mixed excitation of the cepstral speech model.

2.1 Cepstral Speech Synthesis

The source-filter model with cepstral parametrization of the vocal tract transfer function is used for speech synthesis. For voiced speech the filter is excited by a combination of an impulse train and high-pass filtered random noise. For unvoiced speech the excitation is formed purely by a random noise source. The transfer function of the vocal tract model is approximated by Padé approximation of the continued fraction expansion of the exponential function [8].

2.2 Spectral Envelope Transformation

During pleasant emotions the larynx and the pharynx are expanded, the vocal tract walls are relaxed, and the mouth corners are retracted upward. The result is falling first formant and raised resonances. For unpleasant emotions the larynx and pharynx are constricted, the vocal tract walls are tensed, and the mouth corners are retracted downward. The result is more high-frequency energy, rising first formant, and falling second and third formants [1]. We can conclude that the first formant and the higher formants of emotional speech shift in opposite directions. For pleasant emotions the first formant shifts to the left, and the higher formants to the right. For unpleasant emotions the opposite situation occurs: the first formant shifts to the right, and the higher formants to the left.

Although the formant frequencies differ to some extent for different languages and their ranges are overlapped [9], the male voice vowel formant areas without overlap can be determined: F1 ≈ 250 ÷ 700 Hz, F2 ≈ 700 ÷ 2000 Hz, F3 ≈ 2000 ÷ 3200 Hz, F4 ≈ 3200 ÷ 4000 Hz [10]. Thus we can use the frequency of 700 Hz as a border frequency for shifting spectral envelope frequencies of male voice in the opposite directions.

For shifting the first formant and the higher formants in the opposite directions we use a smooth function of frequency $\gamma(f)$ representing formant ratio between emotional and neutral speech - see Fig. 1.

For better analytic representation of this function the frequency scale is logarithmically warped so that 700 Hz corresponds to one fourth of the sampling frequency (4 kHz for 16-kHz sampling) - see Fig. 2.

Analytic expression of the logarithmic frequency scale transformation can be derived from its inverse exponential function

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

where f represents the input frequency and f_t corresponds to the transformed frequency. Three unknown variables a, b, c are determined using three points $[f_t, f]$ = [0 kHz, 0 kHz], [4 kHz, 0.7 kHz], [8 kHz, 8 kHz], as zero frequency and half the sampling frequency remain unchanged. The solution of the system of the three equations is

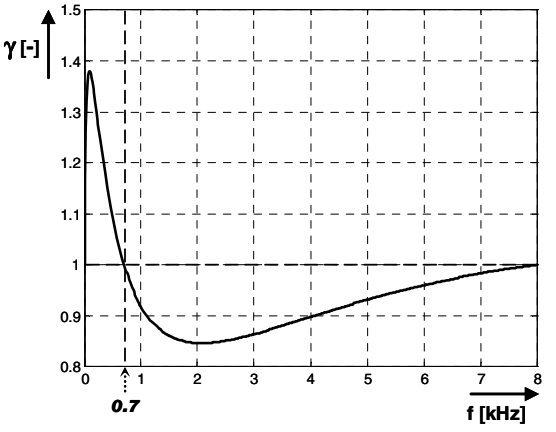

Fig. 1. Formant ratio γ between unpleasant and neutral speech with formant shift to the right for frequencies lower than 700 Hz, and formant shift to the left for frequencies higher than 700 Hz

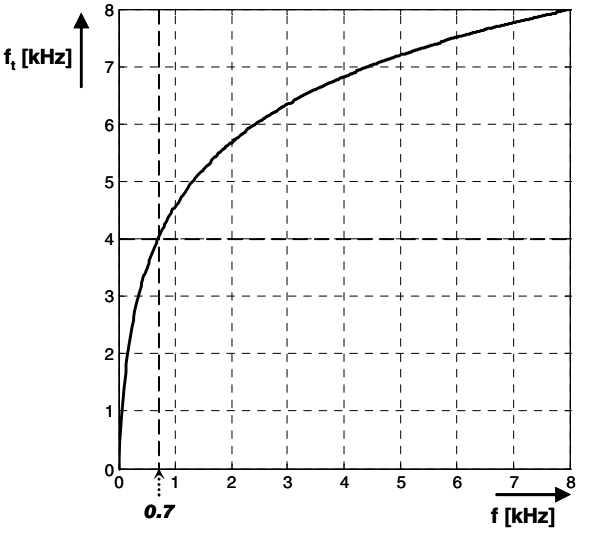

Fig. 2. Logarithmic transformation for frequency scale warping with 700 Hz corresponding to one fourth of the sampling frequency

$$a = \frac{0.49}{6.6}, \qquad b = \sqrt[4]{\frac{7.3}{0.7}}, \qquad c = -a. \tag{2}$$

The transformed frequency is expressed by

$$f_t(f) = \log_b \frac{f + a}{a} = \frac{\ln\left(\frac{f}{a} + 1\right)}{\ln b}. \tag{3}$$

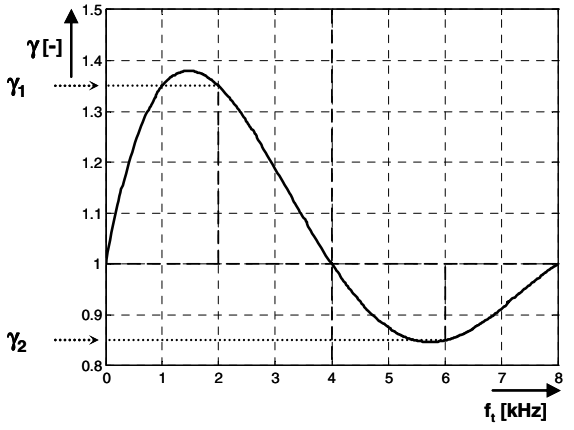

Fig. 3. Formant ratio between unpleasant and neutral speech as a function of logarithmically warped frequency

Formant ratio $\gamma(f_t)$ as a smooth function of the logarithmically warped frequency can be expressed by a fourth-order polynomial function - see Fig. 3

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

Coefficients of this polynomial can be computed from five equidistant points $[f_t, \gamma]$ = [0 kHz, 1], [2 kHz, γ_1], [4 kHz, 1], [6 kHz, γ_2], [8 kHz, 1]. The solution of the system of five equations is

$$p = \frac{1}{48} - \frac{1}{96}\gamma_1 - \frac{1}{96}\gamma_2, \qquad q = -\frac{1}{3} + \frac{3}{16}\gamma_1 + \frac{7}{48}\gamma_2,$$
$$r = \frac{5}{3} - \frac{13}{12}\gamma_1 - \frac{7}{12}\gamma_2, \qquad s = -\frac{8}{3} + 2\gamma_1 + \frac{2}{3}\gamma_2, \qquad t = 1. \tag{5}$$

The modified spectral envelope $E'(f)$ can be computed from the original spectral envelope $E(f)$ using similar relation as used for linear formant modification [11]

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

Using the warping logarithmic function (3), the transformed frequency of 2 kHz corresponds to 165.5 Hz, and 6 kHz corresponds to 2426 Hz. Emotional-to-neutral formant ratios at these frequencies are chosen as shown in Table 1.

Table 1. Chosen emotional-to-neutral formant ratios γ_1 at 165.5 Hz, γ_2 at 2426 Hz

	γ_1	γ_2
joyous-to-neutral	0.70	1.05
angry-to-neutral	1.35	0.85
sad-to-neutral	1.10	0.90

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 9 %, the second and third formants are shifted to the right by about 5 % and the shift gradually decreases. For anger the first formant is shifted to the right by about 12 %, the higher formants are shifted to the left by about 11 % 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 emotional-to-neutral formant ratios in the four formant areas for chosen γ_1, γ_2

	250÷700 Hz	700÷2000 Hz	2000÷3200 Hz	3200÷4000 Hz
joyous-to-neutral	0.9046	1.0589	1.0436	1.0092
angry-to-neutral	1.1215	0.8884	0.8550	0.8842
sad-to-neutral	1.0410	0.9414	0.8999	0.9014

2.3 Spectral Energy Distribution

Different emotions are also characterized by different spectral energy distribution. However, the results of emotional voice analysis are sometimes contradictory. Pleasant emotional speech should have more low-frequency energy; unpleasant one should have more high-frequency energy [1]. On the other hand, increase in high frequencies was reported for joy and anger, and decrease for sadness [12]. It agrees more with our experiments, so we use high-pass filtering for joy and anger, and low-pass filtering for sadness. Both high-pass and low-pass filters are proposed by a standard procedure of second-order shelving filter design [13]. The quality factor is chosen 0.7, and the filter cut-off frequency dividing the frequency range into low and high parts is chosen 1 kHz. The filter gain for joy is chosen 2 dB, for anger and sadness it is chosen 4 dB.

2.4 Spectral Noise

Sad speech should also be accompanied by increased spectral noise [1]. We control its amount by spectral flatness measure according to which the high frequency noise is mixed in voiced frames during cepstral speech synthesis. Its value is chosen 2.5 times higher than that for neutral voice.

3 Prosodic Parameters Modification

Prosody of joyous and angry speech is marked by decreased duration, and increased F0 mean, F0 range, and energy. Opposite trend is observed for sadness where the duration is increased and F0 mean, F0 range, and energy are decreased [1]. However, some authors give different results, e.g., F0 mean and range increased more for sadness than for anger [14]; the same F0 mean for anger and sadness [15]; or too low energy for sadness [16]. We have chosen the results published by Cabral and Oliveira [17] which are in accordance with general information of Scherer [1]. According to [17] the ratio between emotional and neutral F0 mean is 1.18 for joy, 1.15 for anger, and 0.84 for sadness; F0 range ratio is 1.3 for joy and anger, 0.62 for sadness; energy ratio 1.3 for joy, 1.7 for anger, 0.5 for sadness; duration ratio 0.81 for joy, 0.84 for anger, and 1.17 for sadness.

4 Listening Test

Subjective evaluation called "Determination of emotion type" was realized by the listening test located on the web page http://www/lef.um.savba.sk/scripts/itstposl2.dll. The listening test program in the form of MS ISAPI/NSAPI DLL script runs on the server PC and communicates with the user within the framework of the HTTP protocol by means of the HTML pages. Every listening test consists of ten evaluation sets selected randomly from the testing corpus which consists of 80 short sentences with duration 0.9 to 3.6 seconds. The sentences were extracted from the Czech and Slovak stories narrated by male professional actors. For each sentence there is a choice from four possibilities: "joy", "sadness", "anger", or "other".

Twenty eight listeners (21 Czechs and Slovaks, 7 people of other nationalities, 6 women and 22 men) took part in the listening test. The partial results in dependence on sex (male / female listeners) are shown in Fig. 4, partial results in dependence on listeners' nationality (Czech and Slovak / other) are shown in Fig. 5, and the summary results are presented in the form of a confusion matrix in Table 3. Best identified is sadness, worst identified is anger. Joy is sometimes confused with anger and vice versa. Sadness and joy are very rarely confused.

It is clear from the comparison of results in the partial groups that the listeners' sex makes no difference (although, both compared groups were not of the same size) while on the other hand in the group of assessors-foreigners the percentage of successful determination is markedly lower. However, the main proportions among the individual emotions (best identified sadness) are kept in the partial groups, too.

Further, evaluation of the successful determination of the emotion type was carried out in the individual sentences from the testing corpus. Partial results for given three emotions are given in Table 4, Table 5, and Table 6 (values in the column "not classified" represent choice "other" in the listening test, "exchanged" corresponds to incorrectly chosen emotion). Table 7 shows mean values for all emotions ("not correct" comprises "not classified" together with "exchanged"). The overall results present some correlation with partial values for individual emotions, however, there exists also quite contradictory evaluation, supported e.g. by the sentence *s11* being the best for "sadness" and at the same time being the worst for "joy" and "anger".

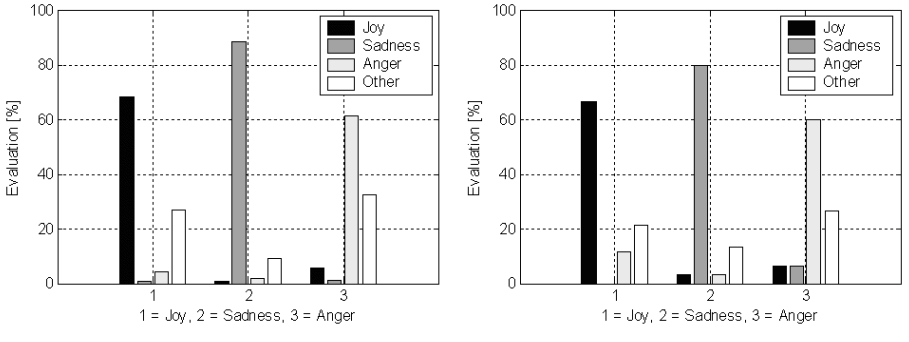

Fig. 4. Partial results of the listening test in dependence on listeners' sex: male (left), female (right)

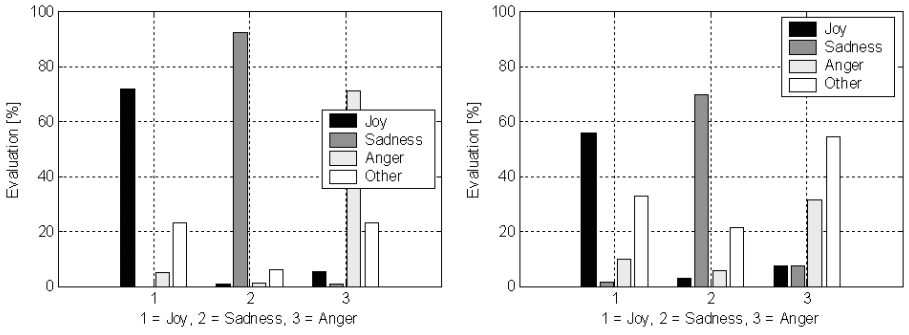

Fig. 5. Partial results of the listening test in dependence on listeners' nationality: Czech and Slovak (left), other nationalities (right)

Table 3. Confusion matrix of the listening test

	joy	anger	sadness	other
joy	**67.86 %**	6.07 %	0.36 %	25.71 %
anger	5.71 %	**61.08 %**	2.14 %	31.07 %
sadness	1.07 %	2.14 %	**86.79 %**	10.00 %

Table 4. Evaluation of the successful determination of the joyous emotion

	sentence	correct	not classified	exchanged
best evaluated	s09*	89 %	11 %	0 %
worst evaluated	s11**	39 %	50 %	11 %

* "Daroval jim kousek úhoru." ("He presented them with a small piece of barrens.")
** "Byl z obyčejného dřeva." ("He was made of a common wood.")

Table 5. Evaluation of the successful determination of the sad emotion

	sentence	correct	not classified	exchanged
best evaluated	s11*	100 %	0 %	0 %
worst evaluated	s12**	71 %	21 %	8 %

* "Byl z obyčejného dřeva." ("He was made of a common wood.")
** "Šel do učení k jednomu truhláři." ("He went to serve his apprenticeship with one joiner.")

Table 6. Evaluation of the successful determination of the angry emotion

	sentence	correct	not classified	exchanged
best evaluated	s14*	80 %	14 %	6 %
worst evaluated	s11**	39 %	61 %	0 %

* "Druhý deň sa konala svadba." ("Next day the wedding took place.")
** "Byl z obyčejného dřeva." ("He was made of a common wood.")

Table 7. Mean values for all emotions

	sentence	Correct	not correct
best evaluated	s14*	84.3 %	15.7 %
worst evaluated	s11**	59.3 %	40.7 %

* "Druhý deň sa konala svadba." ("Next day the wedding took place.")
** "Byl z obyčejného dřeva." ("He was made of a common wood.")

5 Conclusion

Results of the present work have shown that emphasis on spectral modification in emotion conversion gives best (worst) results for sad (angry) speech synthesis which was worst (best) identified in our previous work [18] where only prosodic modification had been done.[1] On the other hand, in the present work, identification of sadness is also improved by lower speech rate, energy, and F0. Joy (positive emotion) and anger (negative emotion) in our speech synthesis differ mainly in the opposite shift of the first formant and the higher formants. It seems that positive and negative emotions are distinguished well, however, instead of "joy" or "anger" many listeners have chosen possibility "other" which implies necessity of wider choice, including fear, disgust, and surprise.

Foreign assessors of the listening test have explained opinion that they did not understand meaning of the utterances and then it was rather difficult for them to decide which emotion they felt when having heard nonsense (from their point of view) sentences. It could be a reason of lower percentage of their successful determination of emotions when compared with the Czech and Slovak listeners.

Contradictory evaluation of the successful determination of the emotion type in some sentences of the testing corpus might be affected also by the fact that we have not always succeeded in selection of suitable sentences uttered in neutral speaking style without any emotion (best pronounced only by the storyteller), but the sentences used in the test were spoken also by different story characters (prince, king, tailor, etc.).

In our next research, we want to experiment with female voice emotion conversion using border frequency between the first and the second formant 840 Hz instead of 700 Hz for male voice, i.e. 20 % higher (Using the general knowledge of [9] that females have on average 20 % higher formant frequencies than males, we had been successful in voice conversion between male and female [19]).

Acknowledgments. The work has been done in the framework of the COST 2102 Action "Cross-Modal Analysis of Verbal and Non-Verbal Communication". It has also been supported by the Ministry of Education of the Slovak Republic

[1] However, quantitative comparison of these two approaches cannot be carried out directly considering the fact that in [18] the speech corpus for the listening test had been created by our text-to-speech system while our new developed method was tested on resynthesis of original speech of actors.

(COST2102/STU/08 – "Audio-Visual Representation of Emotional States") and the Ministry of Education, Youth, and Sports of the Czech Republic (OC08010 – "Emotional Speech Style Analysis, Modelling, and Synthesis").

The authors would also like to express thanks to all the people who participated in the listening test.

References

1. Scherer, K.R.: Vocal Communication of Emotion: A Review of Research Paradigms. Speech Communication 40, 227–256 (2003)
2. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech Emotion Recognition Using Hidden Markov Models. Speech Communication 41, 603–623 (2003)
3. Ververidis, D., Kotropoulos, C.: Emotional Speech Recognition: Resources, Features, and Methods. Speech Communication 48, 1162–1181 (2006)
4. Shami, M., Verhelst, W.: An Evaluation of the Robustness of Existing Supervised Machine Learning Approaches to the Classification of Emotions in Speech. Speech Communication 49, 201–212 (2007)
5. 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.) Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction. LNCS (LNAI), vol. 5042, pp. 213–224. Springer, Heidelberg (2008)
6. Murray, I.R., Arnott, J.L.: Applying an Analysis of Acted Vocal Emotions to Improve the Simulation of Synthetic Speech. Computer Speech and Language 22, 107–129 (2008)
7. Bänziger, T., Scherer, K.R.: The Role of Intonation in Emotional Expressions. Speech Communication 46, 252–267 (2005)
8. Vích, R.: Cepstral Speech Model, Padé Approximation, Excitation, and Gain Matching in Cepstral Speech Synthesis. In: Proceedings of Biosignal, Brno, pp. 77–82 (2000)
9. Fant, G.: Acoustical Analysis of Speech. In: Crocker, M.J. (ed.) Encyclopedia of Acoustics, pp. 1589–1598. John Wiley & Sons, Chichester (1997)
10. Fant, G.: Speech Acoustics and Phonetics. Kluwer Academic Publishers, Dordrecht (2004)
11. Laroche, J.: Time and Pitch Scale Modification of Audio Signals. In: Kahrs, M., Brandenburg, K. (eds.) Applications of Digital Signal Processing to Audio and Acoustics, pp. 279–309. Kluwer Academic Publishers, Dordrecht (2001)
12. Morrison, D., Wang, R., De Silva, L.C.: Ensemble Methods for Spoken Emotion Recognition in Call-Centres. Speech Communication 49, 98–112 (2007)
13. Dutilleux, P., Zölzer, U.: Filters. In: Zölzer, U. (ed.) DAFX – Digital Audio Effects, pp. 31–62. John Wiley & Sons, Chichester (2002)
14. 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)
15. Hirose, K., Sato, K., Asano, Y., Minematsu, N.: Synthesis of F0 Contours Using Generation Process Model Parameters Predicted form Unlabeled Corpora: Application to Emotional Speech Synthesis. Speech Communication 46, 385–404 (2005)
16. 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)

17. Cabral, J.P., Oliveira, L.C.: EmoVoice: A System to Generate Emotions in Speech. In: Proceedings of Interspeech – ICSLP. pp. 1798–1801. Pittsburgh (2006)
18. Přibil, J., Přibilová, A.: Emotional Style Conversion in the TTS System with Cepstral Description. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) COST Action 2102. LNCS (LNAI), vol. 4775, pp. 65–73. Springer, Heidelberg (2007)
19. Přibilová, A., Přibil, J.: Non-Linear Frequency Scale Mapping for Voice Conversion in Text-to-Speech System with Cepstral Description. Speech Communication 48, 1691–1703 (2006)