IDIAP RESEARCH REPORT

SYLLABIC PITCH TUNING FOR NEUTRAL-TO-EMOTIONAL VOICE CONVERSION

Lakshmi Saheer Xingyu Na [a] Milos Cernak

Idiap-RR-31-2015

OCTOBER 2015

[a] Idiap Research Institute

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Syllabic Pitch Tuning for Neutral-to-Emotional Voice Conversion

Lakshmi Saheer[1], Xingyu Na[2], Milos Cernak[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
lakshmi.saheer@idiap.ch,naxingyu@hccl.ioa.ac.cn,milos.cernak@idiap.ch

Abstract. Prosody plays an important role in neutral-to-emotional voice conversion. Prosodic features like pitch are usually estimated and altered at a segmental level based on short windowing of speech signal (where the signal is expected to be quasi-stationary). This results in a frame-wise change of acoustical parameters for synthesizing emotionalized speech. In order to convert a neutral speech to an emotional speech from the same user, it might be better to tune the pitch trajectory at the supra-segmental level like at the syllable-level since the changes in the signal are more subtle and smooth. In this paper we aim to show that the pitch tuning in a neutral-to-emotional voice conversion system may result in a better speech quality output if the tuning is performed at the supra-segmental (syllable) level rather than at frame-level. Subjective evaluation results are shown to demonstrate the improvements in terms of naturalness and speaker similarity.

Index Terms: voice conversion, pitch conversion, syllabic changes, neutral-to-emotional speech conversion

1 Introduction

Speech prosody is the perceived acoustic event with encoded features, such as the emotional state, irony, emphasis, etc. The absence of prosody in writing occasionally result in misunderstanding by the reader. For instance, the acoustical attributes of emotional speech is speaker-dependent. The conversion from neutral to emotionalized speech of the same speaker can be learned by analyzing the pre-recorded speech data [1].

A straightforward way for signal conversion is to scale the samples. Yang et.al. [2] built a high quality real time voice conversion system to alter the speaker identity of a given utterance. However, some of the prior work on pitch conversion [3] show that the modification of pitch is more advanced than simple sample scaling. It was even argued that the variance of the pitch should also be modified independently rather than just deterministically shifting the mean of the pitch [4].

Recent trends in hidden Markov model based statistical speech synthesis could make use of the appropriate data from the speaker to generate expressive

or emotional speech [5]. The basic assumption is that there is enough samples for each emotion from all target speakers. There has been recent efforts to transfer the transformations learned from the expressive data of one speaker to automatically modify the speech characteristics of another speaker [6], meaning that the transformations can be general.

The voice conversion techniques have been developed to convert neutral speech to emotional speech. A few of these techniques raised similar arguments to modify the pitch trajectory at the supra-segmental level [7]. In the research of emotional speech synthesis, the quantitative definition of the acoustic attributes of emotion is necessary for manipulating the output. Rather then defining discrete emotion tag, Tao et.al. [8] classified emotional strength by "strong", "medium" and "weak" to simulate a continuous emotion shift in normal life. The results implied that the emotionalization of speech results in more subtle prosodic pattern changes.

There are two different aspects of prosodic typology: the prominence and the rhythmic patterns of an utterance [9]. While the prominence can be interpreted in the neutral-to-emotional voice conversion as changing the pitch amplitude, still "synchronizing" the pitch with the rhythmical patterns suggests more natural conversion. Thus, proposing pitch changes at syllable level is an simplistic realization of satisfying both the prominence changes at the rhythmical patterns. Our recent work on parametric coding [10] demonstrated the importance of the syllable context prosody encoding. We showed building of the prosody coder that employs the syllable-based pitch, stress and accent parametrization. The previous work thus confirmed that the pitch parametrization at a syllable level is perceived well by the listeners, and in addition, allows to design incremental (online) systems with a syllable latency.

In this paper, we build a neutral-to-emotional voice conversion system and compare the manipulation methods of pitch using sample scaling and syllabic contour modification. The former changes the pitch of the neutral speech independently at the frame level, while the latter tunes the pitch contour at the syllable-level. There is no statistical modeling involved in this work. Deterministic transformations are trained and implemented at the speech signal level. The methods are explained in the next sections followed by the experiments and results.

2 Neutral-to-emotional voice conversion

A neutral-to-emotional voice conversion technique transforms a neutral speech to an affective speech by modifying the relevant speech characteristics. Both prosodic and voice quality features need to be tuned in such a system. In addition, the previous works (e.g. [11]) found that the NLP features may also play a very important role. For example, following rules may be applied in a rule-based emotion conversion system:

– Anger – the highest pitch values is placed on the first content word. The end of the utterances shows strong downward inflections.

– Happiness – the general form of the pitch contour is with a second rising part towards the end of the utterance.

The work of Murray et.al. [12, 13] describe that emotions arise suddenly in response to a particular stimuli, and lasts for seconds or minutes. Murray concludes that the pitch of the voice is important in emotion expression, and voice quality is more important in differentiating discrete emotions. The main aim of this work is to demonstrate that the supra-segmental modification of speech features especially pitch brings in a smooth transition of the speech signal. Pitch, intensity, and speaking rate (in terms of syllabic duration) modification are performed in order to convert a neutral speech to an emotionalized one. The voice quality features are not tuned in this work.

According to the work of Govind et.al. [14], a region wise modification of speech features could broadly incorporate the rule based NLP features similar to the ones mentioned above. This paper also focuses on a region-wise modification for neutral-to-emotional speech conversion. A region-wise feature analysis is performed and the resulting tuning factors are used to transform the corresponding prosodic features. Our work differs from [14] in that the slope of supra-segmental pitch contour is estimated and converted. Also, the pitch parametrization is totally different from the work of Govind et.al.[14]. The following sections present the details on pitch parametrization and the different transformation systems compared in this work. The systems compared here differ only on the pitch transformation.

2.1 Pitch contour parametrization

The pitch value of speech is usually analysed frame by frame. Although the voicing decision breaks pitch trajectory into consecutive pieces, the form of the contour represent the contiguous changes of prosody. To analyse these changes, pitch contour can be parametrized using contour fitting techniques like the Legendre approximations [15]. Cernak et.al. [16] showed that the parametrization of syllabic pitch contours using Legendre approximations is capable of capturing the supra-segmental information. Same Legendre approximations are used in this work. The pitch contour of a syllable of N frames is approximated using the Legendre polynomials as

$$\hat{p}\left(\frac{i}{N}\right) = \sum_{j=0}^{J-1} a_j \cdot \phi_j\left(\frac{i}{N}\right), \quad 0 \leq i \leq N \tag{1}$$

where the parameters a_j are estimated using the pitch values as

$$a_j = \frac{1}{N+1} \sum_{i=0}^{N} p\left(\frac{i}{N}\right) \cdot \phi_j\left(\frac{i}{N}\right), \quad 0 \leq j \leq J-1 \tag{2}$$

and J represents the order of approximation. p is the original pitch trajectory and $\hat{p}$ is the approximated one. In this work, only a second order Legendre function is

used resulting in $J = 2$. The first Legendre parameter a_0 represents the mean of the contour since $\phi_0 = 1$. The second Legendre polynomial is a line. The second parameter a_1, which means the contribution of the line component in the pitch contour, is used to represent the slope. Thus, the pitch of an utterance of Y syllables is parametrized using $2Y$ parameters.

2.2 Pitch contour tuning rules

Pitch tuning is done by applying tuning factors to the pitch of neutral speech. In the frame-wise conversion system, pitch tuning factors are analysed by calculating the ratio of the mean pitch values of emotional speech to that of the neutral speech accordingly [14].

$$\theta_p^{value} = \bar{p}_{emotional}/\bar{p}_{neutral} \tag{3}$$

According to Equation (2), the first Legendre parameter is the mean of the contour. Therefore, given the pitch tuning factors analysed by previous work, such as using the region-wise method in [14], θ_p^{value} is used to tune the first Legendre parameter a_0.

Besides pitch values, the pitch range also varies from neutral to emotional speech. The pitch range of the i^{th} syllable is

$$\Psi_i = \frac{1}{N_i} \sum_{n=1}^{N_i} (p_i(n) - p_{med,i}) \tag{4}$$

where N_i is the syllabic length, and $p_{med,i}$ is the medium pitch value of the syllable. The tuning factor of pitch range is

$$\theta_p^{range} = \Psi_{emotional}/\Psi_{neutral} \tag{5}$$

Tuning of pitch range is done by $\hat{\Psi}_i = \Psi_i \cdot \theta_p^{range}$. We assume that the syllabic pitch range changes proportionally to the slope of the pitch contour. Therefore, θ_p^{range} is also used as the tuning factor for slope, a.k.a. the Legendre parameter a_1.

2.3 The conversion procedure

Once the factors are calculated, a random neutral speech can be converted to an emotionalized speech by altering the corresponding features in a segmental or a supra-segmental flavor. Although the factors are calculated region-wise, they can be used in either shorter or longer units because they describe the average change within a region.

Instead of modifying the pitch values directly, we propose to tune the Legendre parameters a_0 and a_1, then regenerate the pitch contour using Equation (1). Emotional speech is then synthesized using pitch and voice quality features. In section 3, we introduce three neural-to-emotional voice conversion systems for evaluating the pitch tuning methods in this work.

3 Systems

This section presents the details of the different segmental and supra-segmental pitch modification systems compared in this work. The same region-wise transformation rules (similar to the work of Govind et.al. [14]) are analyzed in all systems. Each utterance is divided into three regions:

1. The leading region θ^{lead} - averaged for two first syllables.
2. The trailing region θ^{tail} - averaged for last two syllables.
3. The middle region θ^{mid} - averaged over the rest of the in-between syllables.

The prosodic tuning factors θ^{value} are calculated for each region separately by dividing the average value of emotional and neutral samples. Only the pitch based prosodic modifications are evaluated on segmental or supra-segmental level in each of the three systems compared in this work. Following sections describe systems for pitch modification compared.

3.1 A system – sys0

Given three changing factors for an utterance, the changing factor of a segment is determined by the interpolation of the two neighbouring factors. In other words, the predefined factors for leading, middle and trailing regions are used as discrete data-points at region borders for one-dimensional piecewise linear interpolation $l_n(\theta_p)$. Specifically, the tuned pitch sequence is:

$$\hat{p}_n = (p_n \cdot l_n(\theta_p) - p_{med,n}) \times \theta_p^{range} + p_{med,n}, \quad 0 \leq n \leq N \tag{6}$$

where $l_n(.)$ is the n^{th} interpolant of pitch tuning factor. $p_{med,n}$ is the medium of $p_n \cdot l_n(\theta_p)$ within the utterance. Therefore, the baseline system uses frame-wise modification of both pitch and pitch range using interpolated region-wise rules. Same methods (only shift, without range) are used for modifying intensity and duration for all the three testing systems so that only pitch modification methods are evaluated.

3.2 A system – sys1

In section 2.1, we introduce the parametrization of pitch contour using Legendre polynomials a_0, indicating the mean, and a_1, indicating the slope. In this system, the tuning factor θ_p^{value} is used to shift the syllabic pitch contour in this system using syllable-wise interpolation of region-wise rules, i.e., $\hat{a}_0 = a_0 \cdot l_i(\theta_p^{value})$, where $l_i(\theta_p^{value})$ is the i^{th} syllabic interpolant of pitch tuning factor. The approximated new pitch using Equation (1) is used to replace $p_n \cdot l_n(\theta_p)$ in Equation (6). Thus the pitch range is tuned using the same method as in **sys0**.

3.3 A system – sys2

In this system, pitch tuning is totally based on the Legendre parameters. The Legendre parameters of neutral speech is tuned for $\hat{a}_0$ as in **sys1**, and $\hat{a}_1 = a_1 \cdot \theta_p^{range}$. The tuning factor θ_p^{range} is global. Tuned pitch is generated by Equation (1) using $\hat{a}_0$ and $\hat{a}_1$. Therefore, the pitch contour is shifted using region-wise rules and tilted using the gross-level range factor.

System **sys0** is purely sample-scaling-based, while system **sys2** is purely contour-parameter-based. In between, system **sys1** represents a hybrid of the two systems assuming that the variance of the distribution of the segmental pitch samples reveals more subtle changes than the mean.

4 Experiments

This section presents the details of the experiments performed to evaluate the performance of supra-segmental changes in pitch transformation compared to the frame-level transformation in a neutral-to-emotional speech conversion system. The three systems presented in section 3 are compared for naturalness and speaker similarity.

4.1 Database

The experiments are performed using the Electromagnetic Articulography (EMA) database [17]. The EMA database includes acted emotional speech of a male (AB) and two female (JN, LS) native speakers of American English. A set of ten sentences were recorded five times each with four different emotions, including neutrality, anger, sadness and happiness. Therefore, each speaker contributes 50 parallel emotional speech samples for each emotion digitalized in 12-bit amplitude resolution with 16kHz sampling rate.

4.2 Training

The neutral-to-emotional voice conversion requires factors to be estimated. The (neutral, emotional) speech pairs are syllabified using the Festival[1] front-end, and factors for neutral-to-emotional speech conversion are calculated for the leading, middle and trailing regions of speech in each utterance. Region-wise pitch and intensity are analyzed using Praat[2]. Following features are estimated for the three regions mentioned above.

- Mean pitch $\bar{p}$ (Hz): The average fundamental frequency of the harmonic part of the speech signal.

[1] Festival toolkit: `http://www.cstr.ed.ac.uk/projects/festival/`
[2] Praat software: `http://www.fon.hum.uva.nl/praat/`

– Mean intensity $\bar{i}$ (dB): The intensity of a sound in air is defined as $10\log_{10}(\frac{1}{TP_0^2}\int dt\, x^2(t))$, where $x(t)$ is the sound pressure in units of Pa (Pascal), T is the duration of the sound, and $P_0 = 2 \times 10^{-5}$ Pa is the auditory threshold pressure.
– Mean duration $\bar{d}$ (ms): The average length of syllable-like units.

The tuning factors are estimated as a ratio of the above mentioned mean values for the emotional to neutral speech in each region. Similar to the pitch modification in Equation (3), the intensity and duration tuning factors, θ_i^{value} and θ_d^{value} respectively are calculated as follows.

$$\theta_i^{value} = \bar{i}_{emotional}/\bar{i}_{neutral} \tag{7}$$

$$\theta_d^{value} = \bar{d}_{emotional}/\bar{d}_{neutral} \tag{8}$$

where, $\bar{i}_{emotional}$ and $\bar{i}_{neutral}$ denote the mean intensity of emotional and neutral speech respectively. $\bar{d}_{emotional}$ and $\bar{d}_{neutral}$ denote the mean duration of the emotional and neutral speech respectively. Apart from the pitch mean value, pitch range is also estimated as per Equation (4). All the three systems compared in this paper uses the same set of tuning factors. Only the type of pitch conversion - segmental or supra-segmental is different in each system.

4.3 Subjective evaluation

Subjective tests are performed for evaluating the three voice conversion systems. It is hypothesized that the syllabic level tuning of pitch parameters can lead to better quality[3] of speech. The perceived speaker identifications are expected to remain unchanged in such emotion conversion systems. Hence subjective evaluations are performed in terms of naturalness and speaker similarity respectively. In order to limit the size of the subjective evaluations, only the samples of AB and LS are used in these tests. The factors that are used are shown in Table 1.

Preference test was used for comparing the three systems in terms of naturalness and speaker similarity. For each emotion of each speaker, one utterance was selected for the evaluation. In the naturalness test, two samples generated randomly for each pair of the systems were played. The subject chose his/her preference in naturalness. In the similarity test, besides the pair of generated samples, an original recording of that speaker with the same text and emotion was played as a reference. The subjects judged which of the testing samples was more similar to the reference. Preference in speaker similarity indicates the systems ability to convert emotion without corrupting the vocal print. 16 subjects from various regions participated in the evaluation and all of them use English as their first or second language. The results of subjective evaluation are shown in Fig 1 and 2 with significance level 0.95.

Table 1. *Parameter tuning factors of AB and LS.*

speaker	AB			LS		
emotion	θ_p^{lead}	θ_p^{mid}	θ_p^{tail}	θ_p^{lead}	θ_p^{mid}	θ_p^{tail}
angry	1.62	1.53	1.51	1.03	1.00	1.07
happy	1.56	1.46	1.75	1.27	1.13	1.27
sad	1.26	1.22	1.23	1.00	0.99	1.05
emotion	θ_i^{lead}	θ_i^{mid}	θ_i^{tail}	θ_i^{lead}	θ_i^{mid}	θ_i^{tail}
angry	1.17	1.22	1.26	1.11	1.10	1.10
happy	1.06	1.12	1.15	1.06	1.06	1.08
sad	0.98	1.03	1.07	0.97	0.97	1.00
emotion	θ_d^{lead}	θ_d^{mid}	θ_d^{tail}	θ_d^{lead}	θ_d^{mid}	θ_d^{tail}
angry	1.16	1.03	1.32	1.02	0.93	1.18
happy	1.19	1.02	1.17	1.01	1.01	1.12
sad	1.40	1.35	1.35	1.00	1.06	1.11

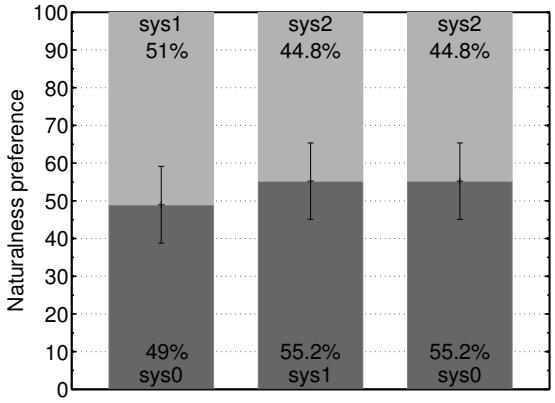


Fig. 1. Preference test of naturalness.

4.4 Analysis

As shown in Fig 1, the naturalness did not show much improvements by using syllabic pitch parameters, but speaker similarity was slightly better according to Fig 2. The **sys0** system is not a purely frame-wise system. We use the region-wise change based on syllable counts as mentioned in [14] which already captures some aspect of supra-segmental change without identifying the type of the segment that is being modified. Even though the differences in performance are not statistically very significant, **sys1** can be attributed as performing slightly better than the other systems in both naturalness and speaker similarity tests. This indicates the importance of supra-segmental modifications of pitch, especially the shift of mean. This proves our hypothesis that supra-segmental transformations

[3] quality here is being quantified using the naturalness score

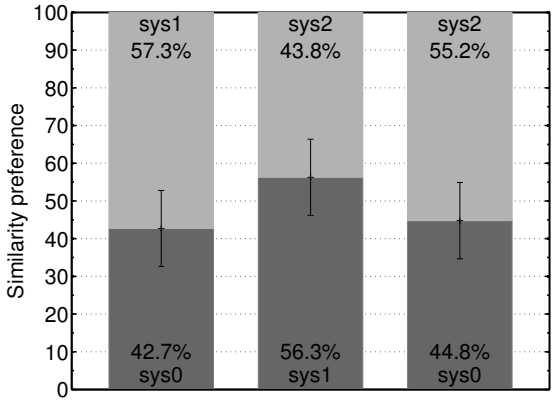


Fig. 2. Preference test of speaker similarity.

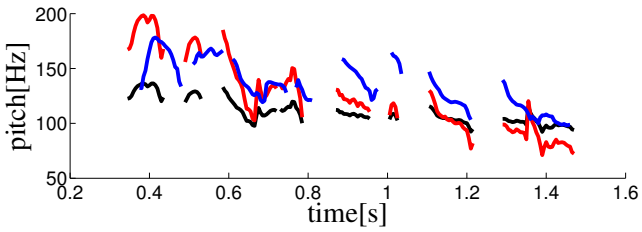
can result in slightly better quality speech. The speaker similarity results are more important indicating that the speaker characteristics are better preserved. There was no statistically significant difference in the emotion recognition subjective evaluation results for the three systems. Hence, those results are not presented here.

An example of the pitch conversion is shown in Fig 3. Pitch of neutral speech, in black, is converted to get an emotional pitch trajectory, in red. The blue pitch trajectory is chosen from the same utterance with the target emotion. It can be observed from 3(a) that, by changing the pitch value and range of each segmental frame, the converted pitch reach a very low pitch value around 70 Hz and jitters, resulting in a loss of both naturalness and speaker similarity. However, by changing Legendre parameters and generating pitch contour from the tuned parameters, the converted pitch trajectories are smoother, and are expressively closer to the target as shown in 3(b) and 3(c). Overall, a smooth transition can be expected from the syllabic pitch tuning.

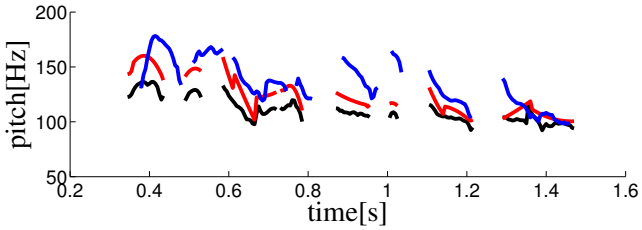
5 Conclusions

This study compared the segmental and supra-segmental neutral-to-emotional speech transformation. Only the pitch modifications were made at these different levels while, the same intensity and duration modifications were performed in all the systems compared. According to the subjective evaluations, the pitch conversion in a neutral-to-emotional voice conversion system does not seem to bring statistically significant improvements in the speech quality output when the conversions are performed at the supra-segmental (syllabic) level. However, the speaker similarity result shows slight improvements. Overall, **sys1** shows better preference scores compared to the other systems.

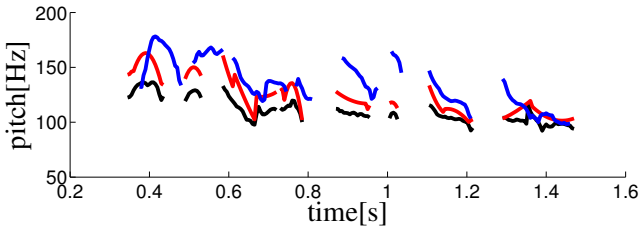
The conversion factors are estimated at the segmental level in the training stage. The same rules/factors as with the segmental system are used for the



(a) Pitch conversion using **sys0**



(b) Pitch conversion using **sys1**



(c) Pitch conversion using **sys2**

Fig. 3. An illustrative sample from speaker AB. Black and blue curves represent the pitch trajectories of neutral speech and the target emotional speech respectively. Red curves are the converted pitch trajectories.

supra-segmental systems **sys1** and **sys2**, which do not actually model or capture any syllable specific rule. The future work is to study whether the syllabic rules should be analysed in a specific manner, such as extending the simplistic approach of the prominence changes at the rhythmical patterns by incorporating the stress and accent prosodic features. In addition, we would like also to investigate alternative supra-segmental F0 representations, such as F0 tilt model, discrete cosine transform and functional data analysis.

6 Acknowledgements

Authors would like to thank the Hasler Geneemo project for funding this research. The work of Xingyu Na was done when he was visiting the Idiap Research Institute and with Beijing Institute of Technology.

References

1. H. Kawanami, Y. Iwami, and T. Toda, "GMM-based voice conversion applied to emotional speech synthesis," in *Proc. of Interspeech*, 2003, pp. 2401–2404.
2. P. F. Yang and Y. Stylianou, "Real time voice alteration based on linear prediction," in *Proc. of ICSLP*, 1998, pp. 1667–1670.
3. D. Chappell and J. Hansen, "Speaker-specific pitch contour modeling and modification," in *Proc. of ICASSP*, 1998, pp. 885–888.
4. T. Ceyssens, W. Verhelst, and P. Wambacq, "On the construction of a pitch conversion system," in *Proc. of EUSIPCO*, 2002, pp. 423–426.
5. J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in hmm-based speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. 88, no. 3, pp. 502–509, 2005.
6. J. Lorenzo Trueba, R. Barra Chicote, J. Yamagishi, O. Watts, and J. M. Montero Martínez, "Towards speaking style transplantation in speech synthesis," in *Proceedings of the 8th Speech Synthesis Workshop*, September 2013, pp. 159–163.
7. M. Wang, M. Wen, K. Hirose, and N. Minematsu, "Emotional voice conversion for Mandarin using tone nucleus model c small corpus and high efficiency," in *Proceedings of Speech Prosody 2012*, 2012, pp. 163–166.
8. J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1145–1154, July 2006.
9. S.-A. Jun, "Prosodic Typology," in *Prosodic Typology: The Phonology of Intonation and Phrasing*, S.-A. Jun, Ed. Oxford University Press, 2005, pp. 430–458.
10. M. Cernak, P. N. Garner, A. Lazaridis, P. Motlicek, and X. Na, "Incremental Syllable-Context Phonetic Vocoding," *IEEE/ACM Trans. on Audio, Speech, and Language Processing (to appear)*, 2015. [Online]. Available: http://publications.idiap.ch/index.php/publications/show/2987
11. I. R. Murray and J. L. Arnott, "Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech," *Computer Speech & Language*, vol. 22, no. 2, pp. 107–129, 2008.
12. ——, "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion," *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, Feb. 1993.
13. ——, "Implementation and testing of a system for producing emotion-by-rule in synthetic speech," *Speech Communication*, vol. 16, no. 4, pp. 369–390, Jun. 1995. [Online]. Available: /idiap/project/emogen/papers/MurrayA05.pdf
14. D. Govind, S. R. M. Prasanna, and B. Yegnanarayana, "Neutral to Target Emotion Conversion Using Source and Suprasegmental Information," in *INTERSPEECH*, 2011, pp. 2969–2972.
15. K.-S. Lee and R. V. Cox, "A very low bit rate speech coder based on a recognition/synthesis paradigm," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 482–491, July 2001.
16. M. Cernak, X. Na, and P. N. Garner, "Syllable-based pitch encoding for low bit rate speech coding with recognition/synthesis architecture," in *Proc. of Interspeech 2013*, Aug. 2013.
17. S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, "An articulatory study of emotional speech production," in *Proc. of Interspeech 2005*, 2005, pp. 497–500.