Intonation Conversion from Neutral to Expressive Speech

Christophe Veaux¹, Xavier Rodet

Ircam, Analysis-Synthesis Team, France

christophe.veaux@ircam.fr, xavier.rodet@ircam.fr

Abstract

Intonation is one of the most important factors of speech expressivity. This paper presents a conversion method for the F0 contours. The F0 segments are represented with discrete cosine transform (DCT) coefficients at the syllable level. Multi-level dynamic features are added to model the temporal correlation between syllables and to constrain the F0 contour at the phrase level. Gaussian mixture models (GMM) are used to map the prosodic features between neutral and expressive speech, and the converted F0 contour is generated under the dynamic features constraints. Experimental evaluation using a database of acted emotional speech shows the effectiveness of the proposed F0 model and conversion method.

Index Terms: prosody conversion, expressivity, F0 modeling, multi-level dynamic features, GMM, DCT

1. Introduction

Emotional voice conversion has mainly been investigated as a post-processing technique to generate expressive speech from a synthesized neutral voice [1,4,5]. However, the applications of emotional voice conversion extend beyond the human-machine interaction, as its principle can be applied to any given natural voice. The work presented here is part of a project² aiming to add emotions to naturally spoken utterances with “neutral” expressivity. Transforming the expressivity in speech requires the manipulation of a wide range of prosodic features such as intonation, intensity, speaking rate, vocal quality and articulation. In this paper, we focus on the modeling and transformation of the F0 contours which constitute the basis of intonation.

The first attempts to transform a neutral intonation into an expressive one were simply based on mean level and range adjustment of the global F0 contour. However, a change in expressivity is likely to affect local F0 patterns, such as the shape and the distribution of F0 prominences, rather than just the global contour. More elaborate transformations of the intonation require, therefore, a parametric model of the F0 contours over the syllables or higher-level segments such as prosodic groups and phrases. Two general approaches can be distinguished here. The first one is a “generative” approach in which context-dependent statistical models of the F0 segments are learned over expressive speech data and used to generate target F0 contours [1]. This generative approach is very similar to the F0 modeling and generation in HMM-based synthesis [2]. It does not exploit the actual F0 contours of the speech signal under transformation and requires knowledge of the textual data. The second approach is inspired by spectral voice conversion techniques [3]. It performs a classification of the F0 parameters and defines a mapping between classes of the source (neutral) and target (expressive) F0 contours. These

classes can be obtained by vector quantization (VQ); or by classification and regression trees (CART) grown after linguistic features [4, 5]; or they can be “soft classes” defined by the components of a Gaussian mixture model (GMM) [5]. The VQ and the GMM techniques do not require any symbolic analysis of the textual data. The idea behind these two techniques is that the clusters defined in the F0 parameter space can correspond to linguistically meaningful patterns such as pitch accent or rising/falling boundaries. In this way, the GMM or VQ clustering can be seen as implicit symbolic analyses performed on the F0 contours. The GMM-based conversion performs better than the codebook-based (VQ) conversion since it does not rely on hard decisions and can be formulated as a mixture of linear regression functions [3].

When transforming a naturally spoken utterance, we believe that it is better to rely on the actual F0 contours to perform the prediction of the converted contours rather than inferring them from the textual data. One problem with the generative approach is that it often results in a stereotyped prosody that does not reflect the natural variability of the F0 contours. However, the classical GMM-based conversion defines only a local mapping between the F0 parameters extracted over a given unit. Therefore, it fails to represent the long-term components of the global F0 contour.

1.1. Proposed method

In this paper, we propose a GMM-based conversion of F0 contours that integrates information about their temporal correlation at multiple levels. The F0 contours over the syllables are represented by their Discrete Cosine Transform (DCT). In addition to this static parameterization, some dynamic features are estimated in order to model the temporal correlations over several levels. At the syllable level, the dynamic features encode the correlation of the shape and mean value between F0 contours. Other dynamic features are used to represent the F0 resetting and the gradient of the mean F0 between phrases. All these dynamic features can be derived from the static parameterization (DCT coefficients) with simple linear transforms. In order to take into account the multi-level dynamic constraints during the conversion, we apply the Maximum Likelihood (ML) parameter trajectory estimation introduced in [6] for spectral conversion. This algorithm simultaneously converts the parameter vectors of all the syllables. In this way, the long-term components of the F0 contour can be modeled and transformed.

1.2. Comparison to other proposals

The proposed F0 model takes some inspiration from the model introduced in [7] in the context of a “generative” approach (prediction of a prosodic contour from the textual data). In this latter model, the F0 contours over the syllables are parameterized with the DCT and dynamic features. Our approach differs, however, in the way it incorporates the higher-level components of the F0 contour. The advantage of using multi-level dynamic constraints to represent these

¹ The author is now at CSTR, Edinburgh University, UK.

² This work was supported by the Respoken FEDER project.

components is that the F0 prediction can be very efficiently determined from a set of linear equations.

A hierarchical prosody conversion is proposed in [8]. The global F0 contour is decomposed in a top-down process where the contours at lower levels are the residuals of the decomposition at the higher levels. A GMM-based conversion is then used for each level. However, if it is acknowledged that the F0 contour has an underlying hierarchical structure, retrieving the actual components at each level is not trivial [9].

2. F0 contour model

The proposed F0 model is defined on a syllable basis. The F0 contours are parameterized over each syllable and the information from the higher levels are projected on the syllable level by extending the static parameterization with specific dynamic features. In our implementation, we use the phrase level in addition to the syllable level. A forced alignment algorithm [10] provides the boundaries of the syllables and phrases.

2.1. Extraction of the F0 contours

The F0 estimates are obtained using the Swipe algorithm [11]. Since there is no valuable information in the unvoiced speech segments, the unvoiced parts are ignored for F0 modeling and conversion. Within each syllable, we define a reliability region over which the F0 contours are selected. The reliability region is given by the local prominence of sonority within the syllable. The sonority profile is estimated by the partial loudness in the low frequency band [300-900] Hz. The maximum of this function corresponds roughly to the nucleus of the syllable. Then starting from this maximum, we extend the borders of the reliability region until a differential threshold of loudness is reached, as illustrated in Figure 1. The F0 contour over this segment is the one considered for modeling and conversion. The syllables whose reliability region has a too short duration are discarded.

2.2. Representation of the F0 contours

Let $s_i = [s_0^{(i)}, \dots, s_{M-1}^{(i)}]$ be a F0 contour of length M extracted over the i th syllable. This contour is represented by its first N DCT coefficients normalized by the factor $1/\sqrt{M}$. The normalized DCT coefficients $[c_0^{(i)}, \dots, c_{N-1}^{(i)}]$ are invariant to the length of the F0 contours and the inverse DCT is defined as:

$$s_m^{(i)} = c_0^{(i)} + \sqrt{2} \sum_{n=1}^{N-1} c_n^{(i)} \cos \left[\frac{\pi}{M} n \left(m + \frac{1}{2} \right) \right]; \quad m = 0, \dots, M-1 \quad (1)$$

The first DCT coefficient is the mean value of the F0 contour. It should also be noticed that the second and third coefficients can be interpreted as the gradient and curvature of the F0 contour. One advantage of this parameterization is that the DCT is an orthogonal transform. Consequently, the mean square error between two linearly time-aligned F0 contours can be simply estimated from the mean square error between their DCT coefficients. Due to the scale-invariance of the DCT coefficients, the duration of the F0 contours must be added to the parameterization. Finally, the F0 contour of the i th syllable is represented by a “static” feature vector of dimension $L = N + 1$:

$$\mathbf{x}_i = [c_0^{(i)}, \dots, c_{N-1}^{(i)}, \log\text{-dur}(i)]^T \quad (2)$$

where $\log\text{-dur}(i)$ denotes the log-duration of the F0 contour.

Figure 1: Extraction of the F0 contour over the local prominence of sonority within each syllable.

2.3. Multi-level dynamic features

The static feature vectors defined in (2) are extended with dynamic features. These dynamic features are used to represent the temporal correlation of the F0 contours between syllables and between phrases. One assumption of our model is that these dynamic features can be used to constrain the long-term components of the global F0 contour.

2.3.1. Syllable level

We introduce three dynamic features at the syllable level:

- (a) $\Delta \mathbf{x}_i$ and $\Delta^2 \mathbf{x}_i$, the delta and delta-delta of the vector $\mathbf{x}_i$;
- (b) $\Delta F0_i^{(beg)}$, the difference between initial F0 value of the current syllable and final F0 value of the previous syllable;
- (c) $\Delta F0_i^{(end)}$, the difference between final F0 value of the current syllable and initial F0 value of the next syllable.

The delta and delta-delta features, respectively $\Delta \mathbf{x}_i$ and $\Delta^2 \mathbf{x}_i$ are estimated according to:

$$\Delta \mathbf{x}_i = \sum_{l=-L_-^{(1)}}^{L_+^{(1)}} w^{(1)}(l) \mathbf{x}_{i+l} \quad (3)$$

$$\Delta^2 \mathbf{x}_i = \sum_{l=-L_-^{(2)}}^{L_+^{(2)}} w^{(2)}(l) \mathbf{x}_{i+l} \quad (4)$$

where $\omega^{(n)}(l)$ are the weights for the delta of order n and $(L_-^{(n)}, L_+^{(n)})$ the estimation range¹. The first coefficients of $\Delta \mathbf{x}_i$

and $\Delta^2 \mathbf{x}_i$ bring some information about the trajectory of the mean F0 over syllables. The higher-order coefficients represent the correlation between the shapes of adjacent syllables. The last coefficients of $\Delta \mathbf{x}_i$ and $\Delta^2 \mathbf{x}_i$ represent the temporal variations of the nucleus durations. These can provide some information about the speech rate as well.

The delta and delta-delta features just add independent constraints to each DCT coefficient although these coefficients are not independent. The inclusion of the two other dynamic features adds some constraints that affect all the DCT coefficients. These features can be easily estimated from the static feature vector $\mathbf{x}_i$. From eq. (1), we can derive:

$$s_0^{(i)} = c_0^{(i)} + \sqrt{2} \sum_{n=1}^{N-1} c_n^{(i)} \cos \left[\frac{\pi}{2M} n \right] \quad (5)$$

$$s_{M-1}^{(i)} = c_0^{(i)} + \sqrt{2} \sum_{n=1}^{N-1} c_n^{(i)} (-1)^n \cos \left[\frac{\pi}{2M} n \right] \quad (6)$$

¹ We choose $L_-^{(0)} = L_+^{(0)} = 0$ and $L_-^{(n)} = L_+^{(n)} = 1$ for $n > 0$, and the weights are: $w^{(0)}(0) = 1$, $w^{(1)}(-1) = -0.5$, $w^{(1)}(0) = 0$, $w^{(1)}(1) = 0.5$, $w^{(2)}(-1) = 1$, $w^{(2)}(0) = -2$, $w^{(2)}(1) = 1$.

It follows that $\Delta F0_i^{(beg)}$ can be expressed as:

$$\Delta F0_i^{(beg)} = s_0^{(i)} - s_{M-1}^{(i-1)} = \mathbf{h}_{(beg)}^T \mathbf{x}_i - \mathbf{h}_{(end)}^T \mathbf{x}_{i-1} \quad (7)$$

where:

$$\mathbf{h}_{(beg)} = [\beta_n^{(beg)}, 0]^T \text{ with } \beta_0^{(beg)} = 1; \beta_{n \geq 1}^{(beg)} = \sqrt{2} \cos\left[\frac{\pi}{2M} n\right]$$

$$\mathbf{h}_{(end)} = [\beta_n^{(end)}, 0]^T \text{ with } \beta_0^{(end)} = 1; \beta_{n \geq 1}^{(end)} = \sqrt{2}(-1)^n \cos\left[\frac{\pi}{2M} n\right]$$

Similarly, $\Delta F0_k^{(end)}$ is defined as:

$$\Delta F0_k^{(end)} = \mathbf{h}_{(end)}^T \mathbf{x}_i - \mathbf{h}_{(beg)}^T \mathbf{x}_{i+1} \quad (8)$$

The inclusion of dynamic features $\Delta F0_i^{(beg)}$ and $\Delta F0_i^{(end)}$ aims to preserve the natural transitions of the F0 contour between syllables. They can also represent some phrase-level information such as the F0 resetting between phrases.

2.3.2. Phrase level

Two dynamic features are introduced at the phrase level: $\Delta R_i^{(beg)}$ the difference of mean F0 level between the current and the previous phrase; $\Delta R_i^{(end)}$ the difference of mean F0 level between the current and the next phrase. These dynamic features can be estimated from the first DCT coefficient of each F0 contour, as it will be shown in the next section. The inclusion of these dynamic features aims to constrain the long-term component of the F0 contour over the sentence.

Finally, we obtain for each syllable i , an extended feature vector of dimensionality $D = 3L + 4$, defined as follows:

$$\mathbf{X}_i = [\mathbf{x}_i; \Delta \mathbf{x}_i; \Delta^2 \mathbf{x}_i; \Delta F0_i^{(beg)}; \Delta F0_i^{(end)}; \Delta R_i^{(beg)}; \Delta R_i^{(end)}] \quad (9)$$

3. Conversion method

We adopt the trajectory-based conversion introduced in [6]. In this approach, the feature vectors of all the syllables over a sentence are simultaneously converted under the dynamic features constraints. It requires the knowledge of the joint distribution of the source and target feature vectors.

3.1. GMM model

For learning the joint distribution, we use parallel corpora of neutral (source) and expressive (target) speech from the same speaker. The syllables of the source and target corpora have been manually labeled and paired together.

Let $\mathbf{Z}_i = [\mathbf{X}_i^T, \mathbf{Y}_i^T]^T$ be the pair of feature vectors of the source $\mathbf{X}_i$ and target $\mathbf{Y}_i$ for a given syllable i . The distribution of the joint vectors $\mathbf{Z}_i$ is modeled with a GMM with K components:

$$p(\mathbf{Z}_i | \lambda) = \sum_{k=1}^K p(k | \lambda) N(\mathbf{Z}_i; \mu_k, \Sigma_k) \quad (10)$$

where $N(\mathbf{Z}_i; \mu_k, \Sigma_k)$ denotes the 2D-dimensional Gaussian with mean $\mu_k = [(\mu_k^{(X)})^T, (\mu_k^{(Y)})^T]^T$ and covariance matrix:

$$\Sigma_k = \begin{bmatrix} \Sigma_k^{(XX)} & \Sigma_k^{(XY)} \\ \Sigma_k^{(YX)} & \Sigma_k^{(YY)} \end{bmatrix} \quad (11)$$

3.2. Prediction of the F0 contours

Let us consider a sequence of source feature vectors $\mathbf{X}_i$ as defined in eq. (9):

$$\mathbf{X} = [\mathbf{X}_1^T, \dots, \mathbf{X}_i^T, \dots, \mathbf{X}_T^T]^T \quad (12)$$

We want to predict the sequence of static feature vectors $\mathbf{y}$ for the target, given the source sequence $\mathbf{X}$. An optimal solution is the Maximum Likelihood (ML) sequence:

$$\hat{\mathbf{y}} = \arg \max P(\mathbf{Y} | \mathbf{X}, \lambda) \quad (13)$$

In the trajectory approach [6], this solution is estimated using an explicit relationship between a sequence $\mathbf{y}$ of static feature vectors and a sequence $\mathbf{Y}$ of extended feature vectors:

$$\mathbf{Y} = \mathbf{W} \mathbf{y} \quad (14)$$

$$\text{with } \mathbf{W} = [\mathbf{W}_1, \dots, \mathbf{W}_i, \dots, \mathbf{W}_T] \text{ and } \mathbf{Y}_i = \mathbf{W}_i \mathbf{y} \quad (15)$$

We can express $\mathbf{W}_i$ according to the components of $\mathbf{Y}_i$:

$$\mathbf{W}_i = [\mathbf{w}_i^{(0)}, \mathbf{w}_i^{(1)}, \mathbf{w}_i^{(2)}, \mathbf{b}_i^{(beg)}, \mathbf{b}_i^{(end)}, \mathbf{d}_i^{(beg)}, \mathbf{d}_i^{(end)}] \quad (16)$$

where $\mathbf{w}_i^{(n)}$ are the weights for the delta of order n :

$$\mathbf{w}_i^{(n)} = [\mathbf{0}_{D \times D}, \dots, \mathbf{0}_{D \times D}, w^{(n)}(-L_-^{(n)}) \mathbf{I}_{D \times D}, \dots, \overbrace{w^{(n)}(0) \mathbf{I}_{D \times D}}^{ith}, \dots, w^{(n)}(L_+^{(n)}) \mathbf{I}_{D \times D}, \mathbf{0}_{D \times D}, \dots, \mathbf{0}_{D \times D}]^T \quad (17)$$

with $\mathbf{I}_{D \times D}$ and $\mathbf{0}_{D \times D}$ the identity and zero matrix of size $D \times D$.

From eq. (7), we have $\Delta F0_i^{(beg)} = \mathbf{b}_i^{(beg)T} \mathbf{y}$ with:

$$\mathbf{b}_i^{(beg)} = [\mathbf{0}_{1 \times D}, \dots, \mathbf{0}_{1 \times D}, -\mathbf{h}_{(end)}^T, \overbrace{\mathbf{h}_{(beg)}^T}^{ith}, \mathbf{0}_{1 \times D}, \dots, \mathbf{0}_{1 \times D}]^T \quad (18)$$

Similarly, from eq. (8), we have $\Delta F0_i^{(end)} = \mathbf{b}_i^{(end)T} \mathbf{y}$ with:

$$\mathbf{b}_i^{(end)} = [\mathbf{0}_{1 \times D}, \dots, \mathbf{0}_{1 \times D}, \mathbf{h}_{(end)}^T, -\mathbf{h}_{(beg)}^T, \mathbf{0}_{1 \times D}, \dots, \mathbf{0}_{1 \times D}]^T \quad (19)$$

Finally, we have $\Delta R_i^{(beg)} = \mathbf{d}_i^{(beg)T} \mathbf{y}$ with:

$$\mathbf{d}_i^{(beg)} = [\mathbf{0}_{1 \times D}, \dots, \mathbf{0}_{1 \times D}, \overbrace{-\alpha_{j-1}^{-1} \mathbf{I}_{1 \times D}, \dots, -\alpha_{j-1}^{-1} \mathbf{I}_{1 \times D}}^{\text{previous phrase } j-1}, \underbrace{\alpha_j^{-1} \mathbf{I}_{1 \times D}, \dots, \alpha_j^{-1} \mathbf{I}_{1 \times D}}_{\text{current phrase } j}, \mathbf{0}_{1 \times D}, \dots, \mathbf{0}_{1 \times D}]^T \quad (20)$$

where α_j is the number of syllables of the phrase j .

The extension to $\Delta R_i^{(end)}$ and $\mathbf{d}_i^{(end)}$ is straightforward.

After [6], the sequence of static feature vectors that maximizes (13) under the constraint of eq. (14) is given by:

$$\hat{\mathbf{y}} = (\mathbf{W}^T \mathbf{D}_m^{(Y)-1} \mathbf{W})^{-1} \mathbf{W}^T \mathbf{D}_m^{(Y)-1} \mathbf{E}_m^{(Y)} \quad (21)$$

where $\hat{\mathbf{m}}$ is the sequence of dominant mixtures, we have:

$$\mathbf{E}_m^{(Y)} = [\mathbf{E}_{\hat{m}_1,1}^{(Y)}, \dots, \mathbf{E}_{\hat{m}_i,i}^{(Y)}, \dots, \mathbf{E}_{\hat{m}_I,I}^{(Y)}] \quad (22)$$

$$\mathbf{D}_m^{(Y)-1} = \text{diag}[\mathbf{D}_{\hat{m}_1,1}^{(Y)-1}, \dots, \mathbf{D}_{\hat{m}_i,i}^{(Y)-1}, \dots, \mathbf{D}_{\hat{m}_I,I}^{(Y)-1}] \quad (23)$$

$$\text{and } \mathbf{E}_{m,i}^{(Y)} = \mu_m^{(Y)} + \Sigma_m^{(YX)} \Sigma_m^{(XX)-1} (\mathbf{X}_i - \mu_m^{(X)}) \quad (24)$$

$$\mathbf{D}_m^{(Y)} = \Sigma_m^{(YY)} - \Sigma_m^{(YX)} \Sigma_m^{(XX)-1} \Sigma_m^{(XY)} \quad (25)$$

Finally, the predicted static feature vectors $\hat{y}$ are converted to F0 contour segments using the inverse DCT defined in eq. (1). The difference between the predicted contours and the source contours defines a transposition pattern. A separate dilatation pattern is derived from the last coefficient of the predicted static vectors (log-duration). All the transform patterns are then applied to the “neutral” speech (source) using a shape-invariant phase vocoder [12].

4. Experiment

A corpus of emotional data was specifically designed for this experiment. A total of 300 sentences including about 5000 syllables were selected. A professional french actor recorded each of these sentences in four expressive styles (Joy, Fear, Sadness, Anger) as well as a neutral style. All utterances were segmented by forced alignment and the segmentation at the syllable level was manually corrected. For this experiment, 250 sentences were used for training, 20 for validation and the rest for testing. The F0 contours were parameterized with 5 DCT coefficients which results in a feature vector of 22 dimensions. For each expressivity, the joint distribution of the “neutral” and “expressive” feature vectors was learnt with an 8-mixture GMM. An example of converted F0 contour in a neutral-to-joy conversion is illustrated in Figure 2.

In order to evaluate the impact of the dynamic features, we estimate the RMSE between target and converted F0 contours for various configurations of the feature vectors. The RMSE is calculated after register normalization for all four expressivities. It can be seen from Table 1 that the addition of the delta and delta-delta bring a dramatic improvement in terms of RMSE. The RMSE is further reduced when we add the other dynamic features.

Features	Static	Static + deltas	All
RMSE (Hz)	31.59	25.85	22.51

Table 1: RMSE between target and converted contours with static, static + deltas, static + multi-level dynamic features.

An informal perceptive test was also conducted with 15 converted speech signals randomly chosen for each expressivity and presented to 10 listeners in mixed orders. Listeners were asked to indicate the perceived emotion of each utterance by selecting one of the four possible emotions. Table 2 illustrates the confusion matrix for the listener’s responses. Fear and Anger are the most confusable, with a large proportion of ‘converted-to-Fear’ utterances rated as Anger. The Sadness has the higher recognition score but still can be confused with Joy. Interestingly, Sadness and Anger are nearly never confused. These two emotions differ mainly on the degree of arousal. On the contrary, Fear and Anger have both a high level of arousal. It seems, therefore, that converted emotions are more easily discriminated along the dimension of arousal than along the dimension of valence (positive/negative). One explanation is that the positive/negative emotions involve the vocal quality (soft/tense) whereas the activation is more correlated with the dynamic range of the F0 contours.

Target	Perceived Emotion			
	Joy	Fear	Sadness	Anger
Joy	64,4	6.8	20,3	8.5
Fear	9.4	55,4	6.4	28.7
Sadness	19,3	6.1	73,4	1.2
Anger	7.1	31.2	1.5	60,1

Table 2: Perceptual Classification Scores.

Figure 2: Conversion of F0 contours from neutral to joy. blue: neutral; red: expressive target; black: converted contours

5. Conclusions

We propose an F0 model that extends the usual set of dynamic features in order to account for the long-term components of prosody. The advantage of using multi-level dynamic constraints to represent aspects of the prosody hierarchy is that the F0 prediction can be very efficiently determined from a set of linear equations. The F0 model represents the contour segments with a discrete cosine transform, which turns out to be an effective and compact parameterization. The conversion is based on the GMM approach with the criteria of Maximum Likelihood Sequence generation. This method is evaluated over a database of acted expressive speech. The objective assessment demonstrates the effectiveness of the model. An informal perceptive test shows that the intonation is not sufficient to generate some expressivities. A complete conversion model should include voice quality.

6. References

- [1] Inanoglu, Z., Young, S., “Intonation Modelling and Adaptation for Emotional Prosody Generation”, ACII, 2005
- [2] Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T., “Modeling of various speaking styles and emotions for HMM Based Speech Synthesis”, Eurospeech, 2003
- [3] Kain, A., Macon, M., “Spectral Voice Conversion for Text-to-Speech Synthesis”, ICASSP, 1998.
- [4] Helander, E., Nurminen, J., “A Novel Method for Pitch Prediction in Voice Conversion”, ICASSP, 2007.
- [5] Tao, J., Yongguo, K., and Li, A. “Prosody Conversion from Neutral Speech to Emotional Speech”, *IEEE Trans. Audio, Speech and Lang Proc.*, vol.14:1145–1153, 2006
- [6] Toda, T., Black, A.W., Tokuda, K., “Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory”, *IEEE Trans. Audio, Speech and Lang Proc.*, 2007
- [7] Latorre, J., Akamine, M., “Multilevel parametric-base F0 model for speech synthesis”, Interspeech 2008.
- [8] Wu, C.H., Hsia, C.C., Lee, C.H., “Hierarchical Prosody Conversion Using Regression-Based Clustering for Emotional Synthesis”, *IEEE Trans. Audio, Speech and Lang Proc.*, 2010.
- [9] J. van Santen and T. Mishra, “Estimating phrase curves in the general superpositional intonation model,” SSW5, 2004.
- [10] P. Lanchantin, X. Rodet, and C. Veaux, “Automatic Phoneme Segmentation with Relaxed Textual Constraints,” LREC 2007.
- [11] Camachao, A., Harris, J.G., “A sawtooth waveform inspired pitch estimator for speech and music,” *JASA*, 124, pp. 1638–1652, 2008.
- [12] Roebel, A., “Shape-invariant speech transformation with the phase vocoder,” Interspeech, 2010.