RISP
NCSP'16
International Workshop on
Nonlinear Circuits, Communications and Signal Processing
Honolulu, Hawaii, USA, March 6-9, 2016

# A study on applying target prediction model to parameterize power envelope of emotional speech

Yawen Xue and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 932-1292, JAPAN
Phone: +81-090-1390-6305
E-mail: xue_yawen, akagi @jaist.ac.jp

**Abstract**

This paper proposes a method to parameterize power envelopes in order to modify the power envelope of neutral speech to convert into emotional speech. Target prediction model that can predict the stable power target in short-term interval is firstly used to estimate the targets of the power envelope. Then we change the targets to a stepwise fucntion by segmenting the starting and ending points in each period. By controling the stepwise targets of the power envelope, a power envelope for emotional speech can be reproduced by 2nd-order critically damped model. Results show that this method show excellent performance for parameterizing power envelopes.

## 1. Introduction

In the field of human-computer-interface (HCI), one of the goals is to improve user experiences by providing genuine human communication. A speech-to-speech translation (S2ST) system [1] plays a consequential role for converting a spoken utterance from one language into another to enable people who speak different languages to communicate. In addition, information conveyed by speech can be divided into three categories: linguistic information, para-linguistic information and non-linguistic information [2]. Linguistic information is considered as the symbolic infomation which is represented by a set of symbols and rules such as the lexical, syntactic and semantic. For para-linguistic information, it is defined as the information that cannot be inferred from the written language but added by the speaker such as intention and attitude. Non-linguistic information shows the information that cannot generally be contolled by the speaker such as age, gender and emotional state. While conventional S2STs focus on processing linguistic information only, which is deficient in synthesizing affective speech, such as emotional rather than neutral speech. Only linguistic information is not enough for people to communicate with each other [3]. Therefore, affective S2STs, that cosider about emotion, [1] is necessary to be taken into consideration.

So far, some previous research considering emotional speech synthesis have already achieved some improvements such as Hidden Markov Model (HMM) with Gaussian Mixture Model (GMM) or concatenative approach, like unit selection [4] [5]. Both methods can synthesize emotional speech with good quality when the emotion is present in a category such as happy, sad, or angry. However, they can only synthesize the emotional speech with the average emotion (not strong or weak emotions) in the emotion category, and both need a huge database for training, although it is difficult to collect many human responses when listening to emotional speech.

In human speech communication, people sometimes strengthen or weaken emotional expressions depending on the situation. Thus, a small number of discrete categories is not sufficient to mimic the emotional speech in daily life. Therefore, some researchers proposed a multi-dimensional approach to express emotion on a continuous-valued scale instead of categorical methods. By using the rule-based synthesis method, tendencies of the variations can be acquired using a small database. With the tendencies of variation, the synthesized speech can convey all degrees of an emotion.

An emotional speech conversion system utilizing a three-layered model for dimensional approach [6] has already proposed by the authors. This system achieved improvement, in which synthesized speech with happy or sad emotions can obtain excellent listening test results comparing with the previous work using two-layered model [7]. For angry speech, however, the synthesized speech cannot give strong impression comparing with happy and sad voice in listening tests. Since power envelope is much related to anger voice, in this paper, we propose a method to parameterize power envelopes of speech using a target prediction model [8] to be able to modify power envelope freely.

The target prediction model can predicts the stable power target in short-term interval by a 2nd-order critically damped system. The target prediction model is firstly used to estimate the target of power envelope. Then the estimated target is changed to a stepwise function. By controlling the magnitudes of the stepwise function of the target, a reproduced power envelope can be predicted using 2nd-order critically

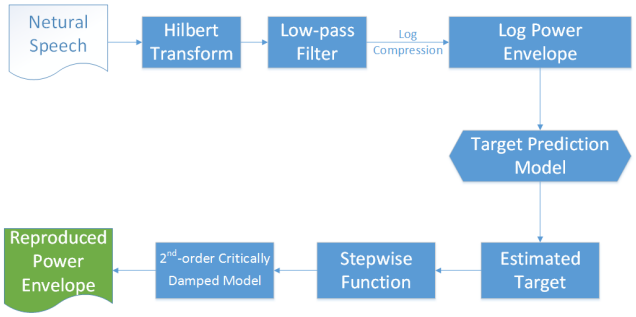


Figure 1: Procedure of reproduing power envelope

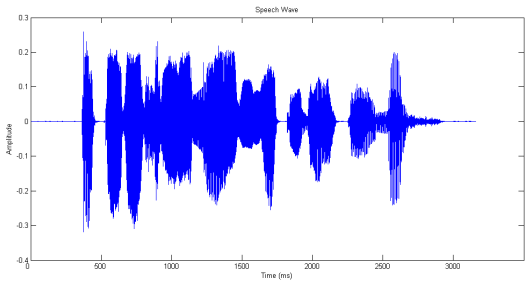


Figure 2: Speech wave of the original speech


Figure 3: Extracted power envelope

damped model that mimics the same process as we model log F0 contour with Fujisaki model in accent control mechanism part [9]. All procedures are shown in Figure 1. In this paper, the original power envelope of neutral speech is firstly extracted and parameterized, then we use the extracted parameter values to get the reproducing power envelope of the same speech. The similarity between the extracted and reproducing power envelopes shows that this method can work well to parameterize power envelope.

This paper is arraged as following. In Section 2, the extraction method of power envelope is shown. In Section 3, target prediction model is conduected to obtain the estimated target of power envelope and then it is changed to the steplike function. In Section 4, the reproducing process of power envelope using 2nd-order critically damped model is discuessed. Section 5 contains a conclusion and a discussion of future work.

## 2. Extracting Power Envelope

The neutral speech signal in Figure 2 is represented as $y(t)$. The power envelpes from $y(t)$ are extracted by

$$e_y(t) = \mathbf{LPF}\left[\left|y(t) + j\mathrm{Hilbert}\left[y(t)\right]\right|^2\right] \tag{1}$$

where LPF[$\cdot$] is a low-pass filtering and Hilbert[$\cdot$] is the Hilbert transform. Then we using Equation 2 to change power envelope in log power envelpe domain.

$$loge_y(t) = 10\log_{10}(e_y(t)) \tag{2}$$

Figure 3 shows the extracted log power envelope.

## 3. Estimating Targets of Power Envelope

A power target prediction model predicts the stable power target in each short-term interval. When the power envelope is approximated by a 2nd-order critically damped system, the model can estiamte target power envelope using short-term power sequences without being given the onset positions of the power trasition.

A 2nd-order critically damped model is generally represented as follows

$$\left(\Delta^2 - 2\lambda\Delta + \lambda^2\right) y_n = \lambda^2 b \tag{3}$$

where $\Delta$ is a differenctial operator in time, $\lambda$ is a reciprocal time constant, time $n = 0$ is the onset position of the transition and $b$ is a target to which $y_n$ converges in the past if $\lambda > 0$ and $n \leq 0$, or in the future if $\lambda < 0$ and $n \geq 0$. The solution of Equation 3 is

$$y_n = (a + cn)\exp(\lambda n) + b \tag{4}$$

where $a$ and $c$ are constant obtained from boundary condition. Previous methods that estimated the parameters of 2nd-order critically damped models have predicted all parameters directly by using Equation 4 and the following measure,

$$e\left(n_0 \text{ or } n_1, \lambda\right) = \sum_{n=n_0}^{n_1} \left|y_n^i - y_n\right|^2, \quad n_0 < n_1 \tag{5}$$

where $y_n^i$ is an unkonwn input sequence. For these methods, a long-term sequence sufficient to start at the onset position of the transition $n_0 = 0$ when $\lambda < 0$ or $n_1 = 0$ when $\lambda > 0$ is essentially required. Then, non-linear optimization under two

values, $n_0$ and $\lambda$ or $n_1$ and $\lambda$ is needed. Howere, the purpose of our target prediction model is to estimate $b$ only.

Divide Equation 3 such that;

$$(\Delta - \lambda)\{(\Delta - \lambda)\, y_n\} = \lambda^2 b \tag{6}$$

and assume that

$$x_n = (\Delta - \lambda)\, y_n \tag{7}$$

$$(\Delta - \lambda)\, x_n = \lambda^2 b \tag{8}$$

By substituting Equation 4 into Equation 7,

$$x_n = c \exp(\lambda n) - \lambda b \tag{9}$$

and Equation 9 is a first-order euqation.

Assuming that

$$c_m = c \exp(\lambda m) \tag{10}$$

at time $n = m$, the neighborhood $x_{m+t}$ of $x_m$ is represented by

$$x_{m+t} = c_m \exp(\lambda t) - \lambda b \tag{11}$$

Thus, if the measure

$$\begin{aligned} e(\lambda) &= \sum_{t=n_0}^{n_1} \left|(\Delta - \lambda) y^i_{m+t} - x_{m+t}\right|^2 \\ &= \sum_{t=n_0}^{n_1} \left|x^i_{m+t} - x_{m+t}\right|^2 \end{aligned}$$

can be used, non-linear optimization under only $\lambda$ is needed and it does not require any knowledge of the onset position of the transition inestimating the target $b$, because $x_{m+t}$ is an exponential fuction. In this prediction, if $\lambda \geq 0$, it is the backward prediction (target in the past). If $\lambda < 0$, it is the forward prediction (target in the future). And we use forward prediction (target in the future) to reproduce power envelope.

In Figure 4, the blue line shows the estimated target of power envelope using the target prediction model.

## 4. Reproducing Power Envelope

The onset point $T_{1j}$ , ending point $T_{2j}$ of every phoneme is segmented manually. After obtaining the estimated power envelope, we extracted the amplitude $A_{aj}$ of the $j$th step in every period of one phoneme. They are inputs of the following fuction that follows the accent mechnism of Fujisaki model. And the stepwise function is shown in Figure 5. Equation 12 shows the step-response. The stepwise input signals to the power control mechanism are difined by their amplitude $Aa$, onset time $T1$ and offset time $T2$ using Euqation

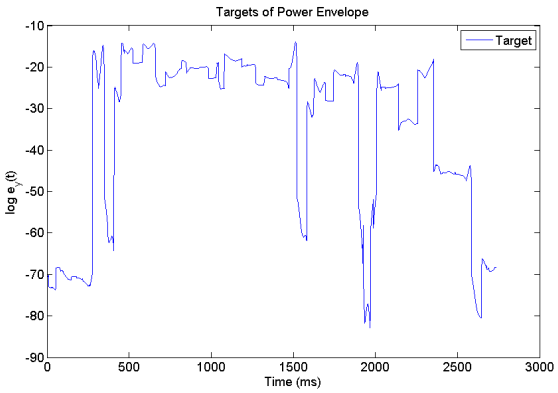


Figure 4: Target of power envelope estimated by target prediction model

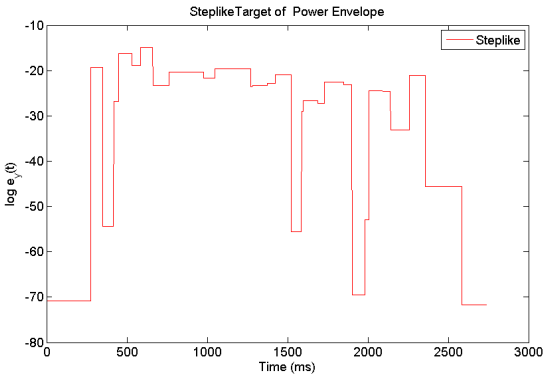


Figure 5: Steplike targets of power envelope

$$Ga_j(t) = 1 - (1 + \delta t)\exp(-\delta t) \quad t \geq 0 \tag{12}$$

where $G_{a(t)}$ represents the step response function. And $loge_y(t)$ is the reproduced power envelope.

$$loge_y(t) = \sum_{j=1}^{J} Aa_j[Ga(t - T1_j) - Ga(t - T2j)] \tag{13}$$

The symbols in these equations forecast

- $Aa_j$: amplitude of the $j$th step, $Aa_j$ is the average value of $b$ in each segmentation,
- $T1_j$: onset of the $j$th step,
- $T2_j$: offset of the $j$th step,
- $\delta$: time constant.

  $\delta$ is the absolute value of the sum of the negative parts of $\lambda$ as we use forward prediction, $\lambda < 0$ (target in the future), to reproduce power envelope.

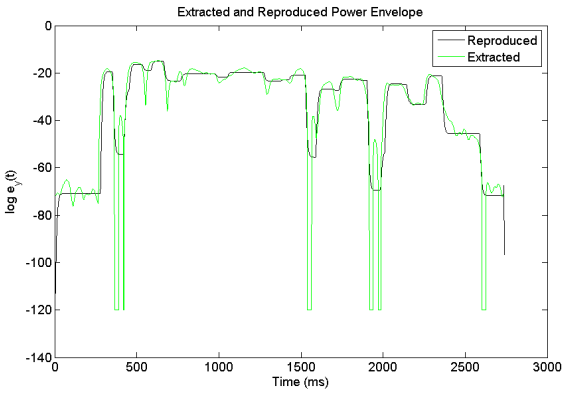


Figure 6: Reproducing power envelope using 2nd-order critically damped model and the extracted power envelope from original speech

In Figure 6, the reproduced power envelope and extracted log power envelope are shown. Signal/Error Ratio (SER) in Euquation 14 and Mean Absolute Error (MAE) in Equation 15 are used to evaluate the difference between the extracted and reproduced power envelope. As the voiced signal is more important of unvoiced part in this research, SER is only calculated in voiced part.

$$SER = 10\log_{10}\frac{\sum_{i=1}^{N}(x_i)^2}{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{14}$$

$$MAE = \frac{\sum_{i=1}^{N}|x_i - y_i|}{N} \tag{15}$$

where $x_i$ is the extracted power envelope and $y_i$ is the reproducing power envelope. $N$ is the number of bit in the voiced part.

The value of SER is 18.01dB and the MAE is about 1.82dB which means that the reproduced power envelope is almost the same as the original extracted power envelope. Therefore, a conclusion can be made that this method can work well for parameterizing power envelope. After this, we modify power envelope by controlling $A_{aj}$ to fit the estimated acoustic features

## 5. Conclusion

In this paper, a method to parameterize power envelope, in which only four parameters ($Aa_j$, $T1_j$, $T2_j$, $\delta$) are used, is proposed. A target prediction model is utilized to esitmate the target of power envelope. The estimated target is changed to steplike function and then the 2nd-order ciritically damped model is conducted to reproduce power envelope. As the SER is large the MAE is small from the reproduced and extracted power envelope, the reproduced power envelope is almost the same as the original extracted power envelope. A conclusion can be made that this method can be used to parameterlize power envelope. In the future, the modifying procedure of power envelope from neutral to emotional speech will be researched later to fit the estimated acoustic features in order to synthesize more human-like emotional speech.

## References

[1] Akagi, M., Han, X., Elbarougy, R., Hamada, Y., & Li, J. “Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages”. Proc. APSIPA2014, CD-ROM, Siem Reap, Cambodia, 2014.

[2] Fujisaki, H. “Information, prosody, and modeling-with emphasis on tonal features of speech”. Proc. Speech Prosody, Nara, Japan, 1-10, 2004.

[3] Elbarougy R., & Akagi, M. “Speech Emotion Recognition System Based on a Dimensional Approach Using a Three-Layered Model”. Proc. of APSIPA CD-ROM, Los Angers, USA, 2012.

[4] Barra-Chicote, R., Yamagishi, J., King, S., Montero, J. M., & Macias-Guarasa, J. “Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech”.Speech Communication, 52(5): 394-404, 2010.

[5] Yamagishi, J., Nose, T., Zen, H., Ling, Z. H., Toda, T., Tokuda, K.. & Renals, S. “Robust speaker-adaptive HMM-based text-to-speech synthesis”. IEEE Transactions on, Audio, Speech, and Language Processing, 17(6): 1208-1230, 2009.

[6] Xue, Y., Hamada, Y.,& Akagi, M. “Emotional speech synthesis system based on a three-layered model using a dimensional approach”. Proc. APSIPA2015, HongKong, 2015.

[7] Hamada, Y., Elbarougy, R., & Akagi, M. “A method for emotional speech synthesis based on the position of emotional state in Valence-Activation space”. Proc. APSIPA2014, CD-ROM, Siem Reap, Cambodia, 2014.

[8] Akagi, M., & Tohkura Y. “Spectrum target prediction model and its application to speech recognition”. Computer Speech and Language .pp.324-344, 1990.

[9] Fujisaki, H., Ohno, S., & Gu, W. “Physiological and physical mechanisms for fundamental frequency control in some tone languages and a command-response model for generation of their f0 contours. In International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages”.