MODELING THE ACOUSTIC CORRELATES OF DIALOG ACT FOR EXPRESSIVE CHINESE TTS SYNTHESIS

Hongwu YANG[*1], Helen M. MENG[2], Lianhong CAI[3]

[*1] College of Physics and Electronic Engineering, Northwest Normal University, 730070 Lanzhou, China
[2] Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, HKSAR, China
[3]Department of Computer Science and Technology, Tsinghua University, 100084 Beijing, China

Keywords: dialog act; expressive text-to-speech synthesis; Pitch Target; GRNN

Abstract

This paper proposed a novel approach for describing the expressivity of dialog text and modelling their acoustic correlates for expressive text-to-speech (TTS) synthesis. We applied the Dialog Acts (DAs) in describing expressivity. In particular, we set up a Wizard-of-Oz (WoZ) data collection framework to collect the tourism domain corpus and annotated the DAs. A Pitch Target model which is optimized to describe Mandarin F0 contours was introduced to model the pitch contour of Mandarin syllables. Then a Generalized Regression Neural Network (GRNN) based model was developed, that can transform acoustic features of neutral speech (parameters of pitch target model, duration, energy and pauses) to resemble expressive speech, according to the DA of the input text. Perceptual evaluation of the modified speech outputs shows that over 63% of the utterances carry appropriate expressivity. Expressive Mean Opinion Score also demonstrated that modified speech improved the expressivity of the neutral speech.

1 Introduction

Expressive speech synthesis has been a hot topic of research in recent years [1]. It has strong potential in enhancing effective communication between human and computers in spoken dialog systems. Expressivity may be a function of the speaker's internal state, the intended effect for the listener, the message content, as well as the state of the dialog. Our recent work focuses on describing and modelling expressivity in relation with the textual content of the message and dialog state for expressive text-to-speech (TTS) synthesis. This leads to two main questions: First, how should we describe the expressivity of the speaker's intended communication (i.e. the message content and dialog state)? Second, how may we render the acoustic features of such intended communication in an expressive way? Previous work has used methods such as categorical annotation scheme [2] and two-dimensional annotation scheme [3] to annotate emotions of the message content. Some efforts have also been made to add emotion and style into the synthesized voice [4-6]. This work focuses on describing expressivity of the message content and modelling their acoustic correlates. Our long term objective is to incorporate expressive TTS for response generation in a spoken dialog system that supports user inquiries in the Hong Kong tourism domain. Our context of study was based on the response messages in a spoken dialog system that belongs to the Hong Kong tourist information domain. There are four main genres of responses messages in this domain: (i) the descriptive genre, where the message describes the attractive features of a scenic spot; (ii) the informative genre, where the message present facts (e.g. opening hours of a tourist spot); (iii) the procedural genre, where the message gives directions (e.g. driving directions); and (iv) the interactive genre, where the message aims to carry forward to the next dialog turn or bring the dialog to a close. The first three message genres have been described in our previous work [7], which relates the message content with the pleasure (P) and arousal (A) descriptors in the PAD emotional model. We proposed that such expressivity should be local to the prosodic word scale and established a nonlinear model to capture the realization of P and A values in terms of expressive acoustic measurements. This study attempts to extend the work in [7] to encompass the interactive message genre. This message genre conveys information regarding the state of the dialog, which may be characterized by the Dialog Act (DA). The DA expresses the communicative goal of a message in the course of a dialog and bears relationships with the neighbouring dialog turns. We propose to relate expressivity with the dialog state (or DA). Different DA may have different expressivity. For example, apologetic utterances may have low dominance level while confirmation utterances may have high dominance level. These two kinds of utterances are expressed with different prosody. In this work, we utilize the DA as a descriptor of expressivity and we aim to model the acoustic correlates of various DA values.

2 Corpora

A Wizard-of-Oz (WoZ) data collection framework was setup to collect the dialog data. 20 dialog acts were used for labelling the dialog sentences. Contrastive neutrals and expressive recordings of the dialog responses were designed to modelling the expressivity with the dialog act.

2.1 Text Corpus of Dialog Responses

The scope of the current study lies in the Hong Kong tourist information domain, which is the backdrop of our spoken dialog system. We use a Wizard-of-Oz (WoZ) data collection setup to elicit interactions in the selected domain from a group of thirty invited subjects. Each subject interacts with a multimodal and multimedia interface, behind which "hides" the wizard. The subjects can issue inquiries using speech, typed text and pen gestures. The wizard can refer to the Discover Hong Kong website during the entire data collection process and always tries to respond to the user's inquiries with best effort. All interactions were logged by the system. The wizard's responses as logged from the WoZ data collection procedure is relatively free form. It contains many disfluencies such as filled pauses, word order reversal due to spontaneity in interactions and tagged information indicating responses in alternative modalities, e.g. highlighted points on a map, urls, etc. In order to ease the subsequent process of modelling the dialog responses, we devised a manual procedure of data regularization where the collected data are simplified into short sentences/utterances with straightforward structures. In total, we have regularized the entire dialog corpus, which consists of 1,500 dialog turns, each with two to five utterances. Overall, there are 3,874 request and response utterances. Table 1 shows a simple dialog example.

Wizard	User
請問，你第二天想去哪里？(Excuse me, where are you going tomorrow?) 請問你想去海洋公園，還是迪士尼樂園？(Would you like to go to Ocean Park or Disneyland?) 這裏是迪士尼樂園的資料，請看。(Here is the information of Disneyland, Please have a look.) 從中環到迪士尼樂園的話，你可以在欣澳站轉乘地鐵迪士尼綫列車就到了。(You can interchange onto the Disneyland Resort Line on Sunny Bay Station if you want to go Disneyland from Central.) 祝你旅途愉快！(Enjoy your trip!)	我想去主題公園看看。(I am planning to go to the popular attractions.) 讓我想想，去迪士尼樂園好了。(I think Disneyland is Ok.) 從中環到這裏怎麼走呢？(How can I get to Disneyland from Central?) 再見。(See you.)

Table 1: An example dialog in Hong Kong tourism domain.

2.2 Annotating the Dialog Act of Text Corpus

The HK tourism domain adopted 20 DAs from VERBMOBIL-2 [8] as shown in Table 2. One utterance corresponds to one dialog act in our corpus. The dialog act of each utterance is annotated by a trained Bayesian Networks (BN) [9]. The labelled DAs also undergo a manual checking for each utterance.

2.3 Speech Corpus of Contrastive Neutral and Expressive Recordings

These contrastive recordings are designed to support our investigation in the acoustic realization of expressive elements in speech. We record contrastive (neutral versus expressive) version of 1,063 selected utterances from the dialog responses mentioned in Section 1.1. These utterances correspond to 6,047 Chinese prosodic words and 13,555 syllables in total. A native Mandarin male speaker was invited to record in a studio. The speaker was asked to record neutral speech with plain and emotionless intonation while to record expressive speech with natural intonation. There are 1,063*2 speech files in total, amounting to over 180 minutes of speech. All recordings were saved in the Microsoft Windows Wav format as sound files (mono-channel, unsigned 16 bit, sampled at 16kHz).

Dialog Act (DA)	Example utterance
CONFIRM	你的票已经订好了(Your ticket has been booked.)
FEEDBACK_POSITIVE	对，这个是带你去看珊瑚(Yes, this brings you to watch coral.)
FEEDBACK_NEGATIVE	没有第二天的票(There are not the other day's tickets.)
CLARIFY	我就是想购物(I would like to go shopping.)
CLOSE	看完了这一个就差不多了(It would be done after this.)
BYE	再见(byebye)
INFORM_DETAILS	可以坐船去(You can take a boat there.)
INFORM_GENERAL	这几个是西贡出名的景点，请看一看(Please have a look at the famous attractions in Sai Kung.)
BACKCHANNEL	好的(Ok.)
OOD	啊(hmm)
COMMIT	要不要我帮你订位？(Can I help you have a seat ?)
DEFER	请等一等(Please wait a minute.)
THANK	非常感谢你的帮忙(Thank you very much for your help)
SUGGEST	第三天去海滩吧！(What about going to beach on the third day ?)
APOLOGY	对不起，我没听清。(Sorry, I beg you pardon?)
REQUEST_COMMENT	这个是坐船吗？(Take a boat?)
REQUEST_PREFERENCE	请问，你第二天想去哪里？(Where would you like to go tomorrow?)
REQUEST_DETAILS	我怎样回到我的酒店呢？(How can I get back to my hotel?)
REQUEST_CLARIFY	请问你收到了没有？(Have you received it ?)
REQUEST_ACTION	小巴的终点是哪里？(Where is the destination of the van?)

Table 2: Twenty DAs and their example utterances.

3 GRNN Based Model for Acoustic Correlates of Expressive Elements

Since the dialog acts (DAs) express the primary communicative function role in the dialog's conversational context, we adopted DAs to be the descriptor for the expressivity of dialog text and used GRNN to model acoustic correlates of expressive elements.

3.1 Acoustic features

Our objective is to capture how expressive elements from transcribed spoken content may be realized in the acoustic speech signal. Acoustic features that are commonly associated with prosody include fundamental frequency (f0), intensity and speaking rate. Therefore we choose to focus on these acoustic features. The f0 contour was modelled with pitch target model. Measurements are taken from the contrastive recordings (neutral versus expressive) of each utterance.

3.2 Modelling syllable's F0 contour with pitch target model

Mandarin is a typical tonal language, in which a syllable with different tone types can represent different morphemes. There are four tone types referred to be "high," "rising," "low," and "falling" [10]. They are mainly manifested by the F0 contours. There have been numerous studies on the tones and intonation of Mandarin. Several quantitative representations have been proposed to describe continuous F0 contours, such as the Fujisaki model [11], the Soft Template Mark-Up Language (STEM-ML) model [12] and the pitch target model [10]. We adopted the pitch target model to modelling the pitch contour of a syllable since the model was originally developed for Mandarin.

Let the syllable boundary be [0,D], the pitch target model is represented by the following equations:

$$\begin{cases} T(t) = at + b \\ y(t) = \beta e^{-\lambda t} + T(t) \end{cases} \quad 0 \le t \le D, \lambda \ge 0 \qquad (1)$$

where $T(\cdot)$ represents the underlying pitch target, and $y(\cdot)$ represents the surface F0 contour. Parameters a and b are the slope and intercept of the underlying pitch target, respectively. These two parameters describe an intended intonational goal by the speaker, which can be very different from the surface F0 contour being observed. Coefficient β is a parameter measuring the distance between F0 contour and the underlying pitch target when t = 0. Parameter λ is a positive number describing how fast the underlying pitch target is approached. Due to the physiological limits of human speech apparatus, these parameters are all subject to the articulatory constraints.

The parameters (a, b, β and λ) can be estimated through nonlinear regression with expected-value parameters at initial and middle points of each syllable's f0 contour. The Levenberg–Marquardt algorithm [13] is used for estimation as a nonlinear regression process.

3.3 Architecture of GRNN based model

GRNN is basis-function architecture that approximates any arbitrary function between input and output vectors directly from training samples. Unlike the conventional multilayer feed-forward neutral network, GRNN is based on nonlinear regression theory for function estimation. Based on Radial-Basis Function network architecture, GRNN needs only a single pass of learning to achieve optimal performance. If the training set consists of vector random variable for x, each with a corresponding value for y, let X be a particular measured value of x, this regression method will produce the estimated value of y, which minimizes the mean-squared error. The conditional mean of y given X can be represented as:

$$\hat{Y}(X) = \left(\int_{-\infty}^{+\infty} y f(x,y)\,dy \right) \Big/ \left(\int_{-\infty}^{+\infty} f(x,y)\,dy \right) \qquad (2)$$

Where, $f(x,y)$ is the joint probability density function of x and y. The function value is estimated optimally as follows:

$$y_i = \left(\sum_{i=1}^{n} h_i w_{ij} \right) \Big/ \left(\sum_{i=1}^{n} h_i \right) \qquad (3)$$

Where, w_{ij} is the target output corresponding to input training vector x_i and output j; $h_i = \exp\left[-D_i^2 / \left(2\sigma^2 \right) \right]$, is the output of a hidden layer neuron; $D_i^2 = \left(X - X_i \right)^T \left(X - X_i \right)$, is the squared distance between the input vector and the training vector; σ is a constant controlling the size of the receptive region.

The architecture of GRNN consist of an input layer, one hidden layer, "un-normalized" output units, a summation unit, and normalized output layer, shown in Figure 1. The hidden and "un-normalized" output units are fully connected. Hidden layer neurons are created to hold the input vector. The weights between the newly created hidden neurons and the output neurons are assigned the target values.

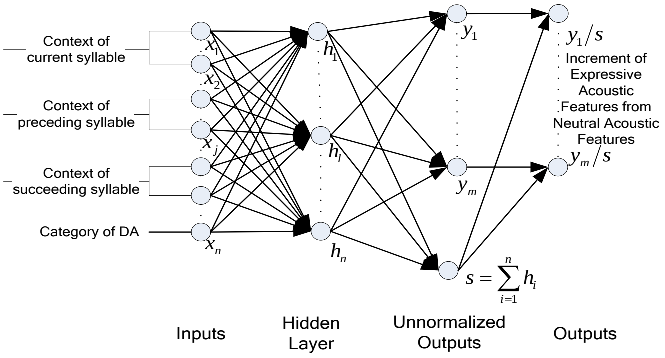
Context of current syllable
Context of preceding syllable
Context of succeeding syllable
Category of DA
x_1
x_2
x_j
x_n
h_1
h_l
h_n
y_1
y_m
$s = \sum_{i=1}^{n} h_i$
y_1/s
Increment of Expressive Acoustic Features from Neutral Acoustic Features
y_m/s
Inputs
Hidden Layer
Unnormalized Outputs
Outputs

Figure 1: Architecture of GRNN.

3.4 Input parameters of GRNN

The GRNN process DAs categories and syllabic level context information as input parameters to generate the percentage increase of syllabic acoustic features in the values of the measurements as one migrates from the neutral speech to its expressive counterpart. The syllable's acoustic features include the pitch target model parameters (a, b, β and λ), duration, energy, as well as pause between syllables. All syllables within an utterance share the DA of the utterance. All syllabic context information is obtained from text analysis, and manually checked the boundary type. The DAs and context information are encoded into discrete values as the input vectors for GRNN. Each syllable has 12 input parameters in total, which may be classified into 6 types shown in Table 3.

Parameter	Description	Value
DA	The DA of current utterance.	1 ~20 for twenty DAs.
Tone	The lexical tone types of the current analyzed syllable, preceding syllable and succeeding syllable.	0 for non, 1 for tone 1, 2 for tone 2, 3 for tone 3, 4 for tone 4, and 5 for light tone.
Pinyin Initial type	Initial types of current syllable and succeeding syllable.	1 for liquids, 2 for fricatives, 3 for vowels, and 4 for nasals.
Pinyin final type	Final types of current syllable and preceding syllable.	1 to 4 for four categories of rhyming vowels (Sihu).
Boundary type	Preceding and succeeding boundary types of current syllable.	1 to 5 for 5 prosodic boundary type.
position	Positions of current syllable in prosodic word and in prosodic phrase.	1 for initial, 2 for middle, and 3 for final

Table 3: Input parameters and their values for GRNN.

3.5 Training the GRNN model

The recorded utterances were automatically segmented into syllables with a home-grown segmentation tool and then the syllable boundaries were checked manually. F0 contour were also checked manually. The pitch target model parameters were estimated from the syllable's f0 contour. Duration，energy and pause were calculated for each syllable. DAs were annotated by a trained BN as mentioned in section 2.3. Contexts of syllables for recorded utterances were obtained by a home-grown text-analysis tool, and also undergo a manual checking of boundary type. There are over 15,000 syllables in corpus. We take 70% of data for the training, and take remained 30% of data for the testing. All training exemplars were normalized to [0,1] with their minimal value and maximal value.

Figure 2 shows an example of pitch, duration and pause transformation using the GRNN based model, in which the neutral pitch contour (top), as well as the duration and pause were transformed into the transformed expressive speech (bottom). The middle figure is the target expressive speech. The utterance is Mandarin sentence "zhe4 liang3 ge4 tuan2 dou1 shi4 zuo4 kuai4 ting3 de5". We can see from the Figure 2 that the transformed pitch contour, duration and pause are close to that of target speech.

4 Evaluation of the Model

We devised a set of preliminary experiments to evaluate the nonlinear model. We selected 100 textual sentences within our tourist information domain and obtain the context using the home grown software tool. We also ensure that the annotated DAs within this set have a good coverage of the distributions in the DA category. We organize a perceptual evaluation whereby each textual sentence is presented to a subject as three speech audio files: (I) a speech recording of neutral speech from the male speaker; (II) a speech recording of expressive speech from the male speaker mentioned above; and (III) a transformation of the speech file from the neutral speech. In this transformation, we use the annotated DAs of each utterance, along with the contexts of each syllable to obtain the predictive increment of acoustic values based on the GRNN model described in Section 3. These parameters are used to transform the speech segment of the corresponding syllable in the neutral waveform. Six of acoustic measurements (all except for pause duration) are modified by the use of STRAIGHT [14]. The pause duration is concatenated to the ends of the syllable in a subsequent step. Transformed speech segments from all the syllables are concatenated in order to form the modified speech utterance with synthetic expressivity, i.e. (III). We invited 13 native speakers of Mandarin to be our subjects in a listening evaluation. For each of the textual sentences, the speech files are played for the subjects in the order of I-II-III or II-I-III. While listening, the subject sees a listing of all the sentence and judges whether a utterance (III) more closely resembles its counterpart in (I) versus that in (II). Results shown that 63% of transformed utterances were been regarded as recorded expressive speech.

We also performed a expressive mean opinion score (EMOS)to evaluate the quality of the transformed speech. The opinion scores (5: excellent expressivity, 4: good expressivity, 3: normal expressivity, 2: worse expressivity, 1: worst expressivity) were given by the subjects after listening the I, II and III. The mean opinion scores were counted after the

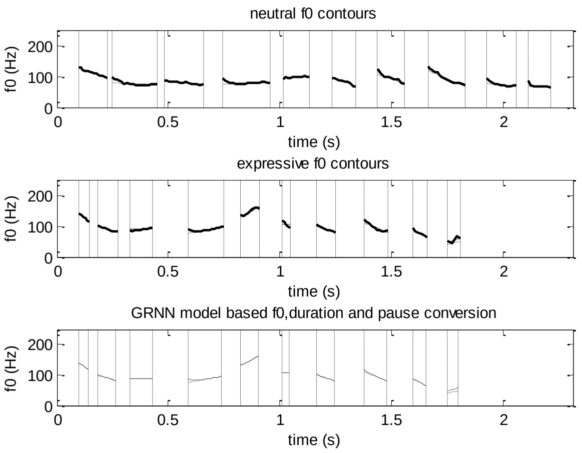

Figure 2: Example of Pitch, duration and pause transformation using GRNN based model from neutral speech to expressive speech. Bold line is the source pitch contour, dash-dot line is the pitch target model generated pitch contour. Vertical dashed line is the boundary of the voiced segment.

testing. The high EMOS represents the higher expressivity. Figure 3 shows the EMOS of the recorded neutral speech, the transformed expressive speech and the recorded expressive speech. We can see from Figure 3 that the EMOS of the transformed speech is high than that of the recorded neutral speech, but is low than that of the recorded expressive speech. This means that the model improved the expressivity of the transformed speech. The voice quality of the transformed speech was degraded due to the modification with the STRAIGHT. This affected the subject's judgement of the EMOS for transformed speech.

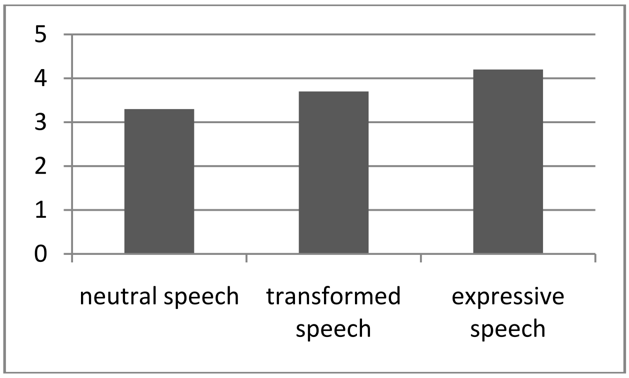

Figure 3: Expressive Mean Opinion Score of neutral speech, transformed speech and expressive speech.

5 Conclusions

This paper proposes a novel approach for describing the expressive elements in interactive text genres and modelling their acoustic correlates for expressive text-to-speech synthesis (TTS). We apply the Dialog Act in describing expressivity. In particular, we found that Dialog Acts are directly applicable in textual content sourced from the tourist information domain. We tokenized the text into syllable and get the contexts of each syllable using a home-grown software tool. We developed a "WoZ" method to collect text corpus. We use the BN to annotate the semantic pairs and manually checked. Analysis of dialog recordings uncovers the acoustic correlates of annotated dialog acts. This enables us to develop a GRNN based model that can transform neutral synthesized speech to become expressive speech, according to the DAs and contexts of the input text. Perceptual evaluation of the speech outputs shows that over 63% of the modified synthesized utterances carry appropriate expressivity. Future work will attempt to extend the model to cover more text genre. The GRNN based model will then be used to enhance the expressivity of our existing Chinese text-to-speech synthesizers for response generation in a spoken dialog system for the tourist information domain.

Acknowledgements

The work is supported by the Key Project of Chinese Ministry of Education (No.208146) and the National Science Foundation of China (NSFC) under grant No. 60433030.

References

[1] N. Campbell. "Towards Synthesizing Expressive Speech: Designing and Collecting Expressive Speech Data", *Proc. Eurospeech*, pp. 1637-1640, (2003).

[2] J. C. Martin, C. Pelachaud, S. Abrilian, L. Devillers, M. Lamolle, M. Mancini. "Levels of Representation in the Annotation of Emotion for the Specification of Expressivity in ECAs", *IVA'05 International Working Conference on Intelligen.*, (2005).

[3] R. Craggs. "Annotating emotion in dialog: issues and approaches", *Proceedings of the 7th Annual CLUK Research Colloquium*, (2004).

[4] J. H. Tao, Y. G. Kang, A. J. Li. "Prosody Conversion From Neutral Speech to Emotional Speech", *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, **volume.** 14, NO. 4, pp. 1145-1154, (2006).

[5] J. Yamagishi, K. Onishi, T. Masuko, T. Kobayashi. "Acoustic Modelling of Speaking Styles and Emotional Expressions in HMM-Based Speech Synthesis", *IEICE Trans. Inf. & Syst.*, **volume** E88-D, No.3, pp. 502-509, (2005).

[6] R. Tsuzuki, H. Zen, K. Tokuda, T. Kitamura, M. Bulut, S. Narayanan. "Constructing Emotional Speech Synthesizers with Limited Speech Database", *Proc. ICSLP*, **volume** 2, pp. 1185-1188 (2004).

[7] H. W. Yang, H. Meng, L. H. Cai. "Modelling the Acoustic Correlates of Expressive Elements in Text Genres for Expressive Text-to-Speech Synthesis", *Proc. Interspeech*, pp. 1806-1809, (2006).

[8] J. Alexandersson, W. Buschbeck, M. K. Fujinami, E. M. Koch, B. S. Reithinger. "Acts in VERBMOBIL-2", Second Edition. Verbmobil Report 226, Universitat Hamburg, DFKI Saarbrucken, Universitat Erlangen, TU Berlin.

[9] H. Meng, W. L. Yip, O. Y. Mok, S. F. Chan. "Natural Language Response Generation in Mixed-Initiative Dialogs using Task Goals and Dialog Acts",. Proc. *Eurospeech* , (2003)

[10] Y. Xu, Q. E. Wang. "pitch targets and their realization: Evidence from mandarin Chinese", *Speech Commun.*, **volume** 33, pp. 319-337 (2001).

[11] H. Fujisaki, K. Hirose. "Analysis of voice fundamental frequency contours for declarative sentence of Japanese", *J. Acoust. Soc. Jpn. (E)*, **volume** 5, no.4, pp. 233-242, (1984).

[12] G. P. Kochanski, C. Shih. "STEM-ML: Language independent prosody description", *Proc. ICSLP, Beijing, China*, pp. 239-242, (2000).

[13] X. Sun "The determination, analysis, and synthesis of fundamental frequency", Ph.D. dissertation, Northwestern Univ., Evanston, IL (2002).

[14] H. Kawahara, J. Estill, O. Fujimura. "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT", *MAVEBA2001, Firentze, Italy*, (2001).