MODELLING THE GLOBAL ACOUSTIC CORRELATES OF EXPRESSIVITY FOR CHINESE TEXT-TO-SPEECH SYNTHESIS

Hongwu Yang[1, 2], Helen M. Meng[2], Zhiyong Wu[2] and Lianhong Cai[1]

[1]Department of Computer Science and Technology, Tsinghua University, 100084 Beijing, China
[2]Department of Systems Engineering and Engineering Management, CUHK, Hong Kong SAR, China

ABSTRACT

This paper proposed a novel approach for describing the expressive elements in dialog response messages for expressive text-to-speech synthesis. We adopt the three-dimensional PAD emotional model in describing expressivity based on response message content and its dialog state. In particular, we use the *P* (pleasure) and *A* (arousal) descriptors to describe expressivity at the local, prosodic-word level based on its semantics. We also use the *D* (dominance) descriptor to describe expressivity at the global, utterance level based on its dialog act. Our context of study is based on response messages of a spoken dialog system in the Hong Kong tourism domain. We also prepared contrastive (neutral versus expressive) recordings to aid identification of the acoustic correlates of expressivity at both local and global levels. We utilized the acoustic analysis of these contrastive recordings to establish a nonlinear model that can be used to modulate input neutral speech at both local and global levels to generate output expressive speech. This work focuses on the nonlinear relationship between the *D* (dominance) values and their acoustic correlates. Perceptual evaluation indicates that local modulation of input neutral speech produces over 73% utterances carry appropriate expressivity. The combined uses of both local and global modulations produce nearly 84% expressive utterances.

Index Terms— Expressive text-to-speech synthesis, non-linear model, dialog act, PAD emotional model

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 modeling 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? Our recent work adopts the PAD emotional model [2] to describe the expressivity of both the message content (in textual form) and the dialog state. The PAD emotional model describes emotional states along three nearly orthogonal dimensions: (i) "pleasure-displeasure" (*P*) distinguishes the positive-negative affective quality of emotional states; (ii) "arousal-nonarousal" (*A*) refers to a combination of physical activity and mental alertness; and (iii) "dominance-submissiveness" (*D*) is defined in terms of control versus lack of control. The implementation of the PAD three-dimensional space uses axes ranging from -1.0 to 1.0 for each dimension. Our context of study is 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 [3], which relates the message content with the pleasure (P) and arousal (A) descriptors in the PAD 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 [3] 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 neighboring dialog turns. We propose to relate expressivity based on the dialog state (or DA) with the D descriptor in the PAD model. Different DA may have different dominance levels. For example, apologetic utterances may have low dominance level (lack of control) while confirmation utterances may have high dominance level. These two kinds of utterances are expressed with different prosody. In this work, we utilize the D (dominance) parameter as a global (i.e. utterance-level) descriptor of expressivity due to the DA and we aim to model the acoustic correlates of various D values.

We reference Chao Yuen-ren's theory [4], which states that the expressive intonation is the combination of "small ripples" (syllabic tone) and "large waves" (intonation). Hence, this work also considers the combination of the local (i.e. within each prosodic word) and global (i.e. for the entire utterance) levels of expressivity covering all four message genres in the tourist information domain. Two response messages with identical textual content will have the same local expressive characteristics but different global characteristics. For example, the response "OK" may confirm user's input statement, while "OK?" may seek to acquire confirmation from the user. The former should be expressed with declarative intonation while the latter should be expressed with interrogative intonation. We propose a two-step sequential process for combining local and global expressivity, as shown in Figure 1: (i) modulate input neutral speech with the

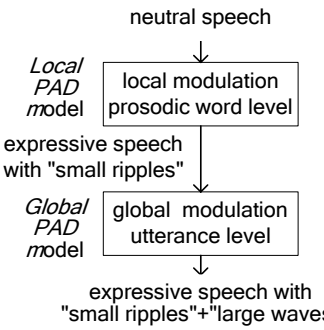

Fig. 1. Framework of Expressive Speech Generation.

local model at the prosodic word level to obtain expressive speech with "small ripples"; and (ii) modulate the input speech with the global model at the utterance level to obtain expressive speech with "large waves".

In the following, we will describe our work in collecting a dialog response corpus as well as recording contrastive utterances based on the responses, determining descriptors of expressivity at the utterance and prosodic word levels, determining acoustic correlates of each descriptor, modeling the correlates, as well as applying the model to modulate neutral speech to become expressive speech. We believe that such an expressive model can be used to modulate the neutral outputs of our TTS systems based on the genres of spoken messages in the domains of interest. The proposed framework can be generalized to achieve expressive speech synthesis for domain-specific spoken dialog systems.

2. CORPORA

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 modeling 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 1500 dialog turns, each with two to five utterances. Overall, there are 3874 request and response utterances.

2.2. 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 2.1. 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 1063*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).

3. ANNOTATING EXPRESSIVITY IN RESPONSES

We adapted twenty dialog acts (*DA*s) based on VERBMOBIL-2 [5] for the current dialog domain in Hong Kong tourism information. Each response utterance corresponds to a dialog act in the corpus. The dialog act of an utterance is annotated automatically by trained Bayesian Networks (BN) [6] and the annotations are then checked manually. The utterances are also automatically segmented into prosodic words by means of a home-grown software tool that applies a set of heuristic rules previously induced from data [7]. We laid down a set of principles for annotating the global expressivity of the dialog act with *D* values and the local expressivity of the prosodic words with *P* and *A* values.

3.1. Principles of Annotation of PAD Values

This subsection describes our principles of assigning values to the PAD emotional model. Assignments are based on the response texts and their related dialog act and semantic concept annotations.

(i) ***D* values:** Utterances that confirm or feedback information are labeled with *D*=1 (for very dominant). Utterances that give introductions, explain facts or bring dialogs to a close are labeled with *D*=0.5 (for dominant). Utterances that give suggestions, express thanks, ask for help or deferment are labeled with *D*=-0.5 (for submissive). Apologetic utterances and interrogative utterances are labeled with *D*=-1 (for very submissive). Other utterances are labeled with *D*=0. Table 1 presents some examples of the correspondences between dialog acts and their *D* values.

(ii) ***P* values:** Commendatory words or words with positive connotations are labeled with *P*=1 (for pleasure). Derogatory words or words with negative connotations are labeled with *P*=-1 (for displeasure). Other words are labeled with *P*=0.

(iii) ***A* values:** Superlatives and words denoting a high degree or extent are labeled with a maximum degree of arousal, i.e. *A*=1. Comparatives and words carrying key facts which should be emphasized (e.g. street names, transportation means, etc.) are labeled with an intermediate degree of arousal, i.e. *A*=0.5. Other words are labeled with *A*=0. A common construct found in our current corpus is "…not only <phrase1>, but also <phrase2>…." We annotate prosodic words in <phrase1> with *A*=0.5 and those in <phrase2> with *A*=1.

3.2. Results of Annotation

As mentioned in the previous subsection, *global* expressivity over the whole utterance is described by the *D* parameter value. The values may be assigned by table lookup based on the annotated *DA*s. Table 2 shows the statistics related to annotated utterances. 34.6% of the utterances are labeled with *D*=-1 values. 26.3% are labeled with *D*=0.5 values, mainly because the utterances in the interactive genre are asking or answering questions. Remaining *D* values all cover over 10% of the utterances.

To describe local expressivity within each utterance, we annotate *P* and *A* values with reference to the semantic concept corresponding to each prosodic word. We begin with regular

Table 1. Twenty DAs used in our corpus , their example utterances and their D values.

Dialog Act (*DA*)	Example utterance	*D* value
CONFIRM FEEDBACK_POSITIVE FEEDBACK_NEGATIVE	你的票已经订好了。 (Your ticket has been booked)	1
CLARIFY, CLOSE, BYE INFORM_DETAILS INFORM_GENERAL	我就是想购物 (I just want to go shopping.)	0.5
BACKCHANNEL,OOD	啊 (hmm…)	0
COMMIT, DEFER THANK, SUGGEST	请等一等 (Please wait a minute.)	-0.5
APOLOGY REQUEST_COMMENT REQUEST_PREFERENCE REQUEST_DETAILS REQUEST_CLARIFY REQUEST_ACTION	对不起，我没听清。 (Sorry, I beg you parden?) 小巴的终点是哪里？ (Where is the destination of the van?)	-1

Table 2. Statistics of annotated *D* values for utterances in dialog corpus.

D	-1	-0.5	0	0.5	1
#utterance (total: 1,063)	351	145	141	316	110
% of occurrence	34.6	14,3	13.9	26.3	10.9

Table 3. Statistics of annotated *P* and *A* values for prosodic words (PW) in dialog corpus.

(P,A)	(0,0)	(0,0.5)	(-1,0.5)	(1,0.5)	(1,1)
#PW (total: 6,047)	3161	2419	66	339	62
% of occurrence	52.3	25.8	1.1	10.4	1.0

expressions for concept tagging. There are 790 semantic concepts in total. The *P* and *A* values of each concept are annotated by two annotators. The concepts with ambiguous labels (~13%) undergo a third pass in annotation to resolve ambiguity. This is followed by table lookup of a mapping between semantic concepts with *P* and *A* values. Hence the prosodic words in an utterance can be automatically labeled with *P* and *A* values. The labels are then checked manually.

Table 3 shows the statistics related to annotated prosodic words. Since we are using a different corpus from [3] in order to include good coverage of all message genres, the results are new. 1.1% of the prosodic words are labeled with negative *P* values, corresponding to a negative answer or apology. No prosodic word was found to carry the annotations of (*P*=0, *A*=1), (*P*=1, *A*=0) and (*P*=-1, A=0). Therefore, this study proceeded with a focus on the five types of (*P*, *A*) combinations (Table 3) for local expressivity.

4. ACOUSTIC ANALYSIS OF GLOBAL EXPRESSIVE FEATURES

Our objective is to analyze 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. We capture both the average and the dynamicity of these acoustic features in six measurements from each utterance:

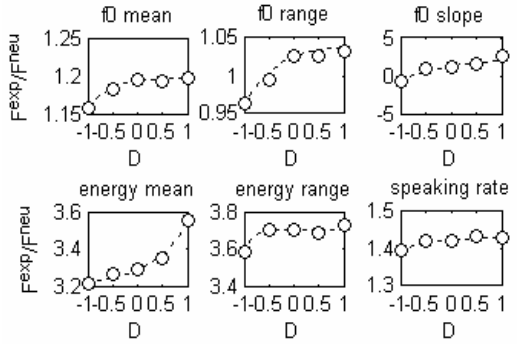

Fig. 2. Difference of expressive features from neutral features for different *D* values. F^{exp}/F^{neu} is the increase of the expressive feature from the neutral feature.

- **Intonation:** f0 mean, f0 range and f0;
- **Intensity:** mean and range of the RMS energy;
- **Speaking rate:** syllables per minute.

Measurements are taken from the contrastive recordings (neutral versus expressive) of each utterance. We also compute the percentage increase in the values of the measurements as one migrates from the neutral speech to its expressive counterpart. Results are shown in Figure 2 with circles (ignore the dotted lines temporarily) for different *D* values. We can see from Figure 2 that there is a general increase in expressive acoustic measurements when the *D* value increases. However, the f0 range and the energy range have a slight decrease at *D*=0.5 compared with *D*=0. The increase of f0 slope has negative value at *D*=-1. This is probably because *D*=-1 describes interrogative intonation (see Table 1), which has a raised pitch contour at the final of utterance. Although the increases are different among the six features, they all have a non-linear relationship with the increase of the *D* values.

5. MODELLING THE GLOBAL ACOUSTIC CORRELATES OF EXPRESSIVE FEATURES

Based on the observations of Figure 2, we propose a nonlinear model to capture the relationship between the *D* values with the acoustic measurements, as shown in Equation (1).

$$\frac{F^{\exp}}{F^{neu}} = C_1 D \exp\left(-C_2 D\right) + C_3 \qquad (1)$$

where F^{exp} is the feature from expressive speech, F^{neu} is the feature from neutral speech, F^{exp}/F^{neu} is the increase of the expressive feature from the neutral feature, and C_1, C_2 and C_3 are coefficients. This equation captures that the relative increase of the expressive measurement bears an exponential relationship with that of the neutral measurement. We use nonlinear least-squares regression to estimate the constant values in Equation (1). In order to produce an accurate finite-difference gradient, the initial coefficients are set to 1 and the maximum number of iterations is 100. We use all utterances to find the coefficients that best fit the model. Results of regression are shown in Figure 2 in dotted lines. We can see from the figure that the outputs of model (dotted lines) track the actual measurements (circles) very well.

6. MODULATION PROCESS

6.1. Local Modulation

Local modulation takes place for each prosodic word unit, based on its annotated (*P, A)* values. The nonlinear model as described

in [3] is used to compute F^{exp}/F^{neu} which is then used to modulate the recorded neutral waveform of the corresponding prosodic work segment. The modulation is realized by the use of STRAIGHT [8]. Pause segments with appropriate durations are concatenated to the starting and ending points of the prosodic segment in a subsequent step. Modulated, expressive speech segments from all the prosodic words are then concatenated in order to generate a synthetic speech utterance that carries local expressivity.

6.2. Global Modulation

Global modulation takes place for an entire utterance based on the D value derived from the dialog act of the utterance. The nonlinear model as described in Equation (1) is used to obtain F^{exp}/F^{neu}, which is then used to modulate the six acoustic measurements of the input neutral utterance. Again, modulation is realized by the use of STRAIGHT [8] to generate the synthetic output utterance that carries global expressivity.

7. EVALUATION

We devised a set of preliminary experiments to evaluate the nonlinear expressive models. We selected 50 textual responses within the tourist information domain. We acquired their annotated *DA*s and also applied a homegrown software tool [6] to extract the constituent prosodic words. We also ensure that the annotated *D* values for utterances and *P, A* value for prosodic words within this set have a good coverage of the distributions in the *PAD* space. We organize a perceptual evaluation whereby each textual response is presented to a subject in terms of five speech audio files: (I) a recording of neutral speech from the male speaker mentioned above; (II) a recording of expressive speech from the same speaker; (III) a locally modulated speech file; (IV) a globally modulated speech file; and (V) a speech file that is first modulated by the local model then modulated by the global model. We invited 14 native speakers of Mandarin to be our subjects in a listening evaluation. For each of the responses, the speech files are played for the subjects in the order of I-II-X or II-I-X. Here, X is the modified speech, e.g. randomly as one of (III), (IV) or (V). While listening, the subject sees a listing of all response texts and judges whether an utterance X more closely resembles its counterpart in (I) versus that in (II). Results are shown in Table 4. We can see that the combined local and global modulation outperforms the local-only and global-only modulations. The result of local modulation is better than that of global modulation, which suggests that the local changes in acoustic features is more important perceptually than the overall increase of average feature values derived from the neutral utterance. Furthermore, the combined use of both local and global modulations gave the best performance. This result seems compatible with Chao Yuen-ren's theory [4].

8. CONCLUSIONS AND FUTURE WORK

This paper proposed a novel approach for describing the expressive elements in dialog response messages for expressive text-to-speech synthesis. We adopt the three-dimensional PAD emotional model in describing expressivity based on response message content and its dialog state. In particular, we use the *P* (pleasure) and *A* (arousal) descriptors to describe expressivity at the local, prosodic-word level based on its semantics. We also use the *D* (dominance) descriptor to describe expressivity at the global, utterance level based on its dialog act. Our context of study is based on response messages of a spoken dialog system in the Hong Kong tourism domain. We also prepared contrastive (neutral versus expressive) recordings to aid identification of the acoustic correlates of expressivity at both local and global levels. We utilized the acoustic analysis of these contrastive recordings to establish a nonlinear model that can be used to modulate input neutral speech at both local and global levels to generate output expressive speech. Perceptual evaluation indicates that local modulation of input neutral speech produces over 73% utterances carry appropriate expressivity. The combined use of both local and global modulations produces nearly 84% expressive utterances. This result seems compatible with Chao Yuen-ren's theory. Future work will attempt to extend the model to cover more (*P, A*) values for global model and *D* values for local model. The model will then be used to enhance the expressivity of our Chinese text-to-speech synthesizers for response generation in a spoken dialog system for the tourist information domain.

Table 4. Results of the listening evaluation. %UT denotes the percentage of modulated utterances judged to bear closer resemblance with its counterpart in expressive version versus that in neutral version. Local modulation means only local model was used to transform the neutral speech. Global modulation means only global model was used to transfor neutral speech. Local+global modulation means both local and global models were used to transform neutral speech.

	local modulation	global modulation	local+global modulation
#UT	512	455	587
%UT	73.1%	65.0%	83.9%

9. ACKNOWLEDGEMENTS

The work is supported by the National Science Foundation of China (NSFC) under grant No. 60433030 and the research fund from the joint fund of NSFC-RGC (Research Grant Council of Hong Kong) under grant No. 60418012 and N-CUHK417/04. The work is affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

10. REFERENCES

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

[2] Mehrabian., A.: Framework for a comprehensive description and measurement of emotional states. Genet Soc Gen Psychol Monogr (1995) 121(3):339-61.

[3] Yang, H.W., Meng, H., Cai, L.H.: Modeling the Acoustic Correlates of Expressive Elements in Text Genres for Expressive Text-to-Speech Synthesis. Proc. Interspeech (2006)

[4] Chao, Yuen-ren.: A grammar of spoken Chinese. Berkeley: University of California Press (1968)

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

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

[7] Zhao, S., Tao, J.H., Cai, L.H.: Prosodic Phrasing with Inductive Learning. Proc. ICSLP (2002)

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