Modeling the Expressivity of Input Text Semantics for Chinese Text-to-Speech Synthesis in a Spoken Dialog System

Zhiyong Wu, Helen M. Meng, *Member, IEEE*, Hongwu Yang, *Associate Member, IEEE*, and Lianhong Cai

***Abstract*—This work focuses on the development of expressive text-to-speech synthesis techniques for a Chinese spoken dialog system, where the expressivity is driven by the message content. We adapt the three-dimensional pleasure-displeasure, arousal-nonarousal and dominance-submissiveness (PAD) model for describing expressivity in input text semantics. The context of our study is based on response messages generated by a spoken dialog system in the tourist information domain. We use the P (pleasure) and A (arousal) dimensions to describe expressivity at the prosodic word level based on *lexical semantics*. The D (dominance) dimension is used to describe expressivity at the utterance level based on *dialog acts*. We analyze contrastive (neutral versus expressive) speech recordings to develop a nonlinear perturbation model that incorporates the PAD values of a response message to transform neutral speech into expressive speech. Two levels of perturbations are implemented—local perturbation at the prosodic word level, as well as global perturbation at the utterance level. Perceptual experiments involving 14 subjects indicate that the proposed approach can significantly enhance expressivity in response generation for a spoken dialog system.**

***Index Terms*—Expressive text-to-speech (TTS) synthesis, nonlinear perturbation model, response generation, spoken dialog system (SDS).**

I. Introduction

THIS work aims to develop an expressive text-to-speech (E-TTS) synthesizer to serve as an integral output channel in a spoken dialog system (SDS). Our long-term goal is to personify the multimodal system output in the form of an avatar that can converse naturally with the user. The interchange should ideally resemble two interlocutors who seek to attain an information goal collaboratively through the course of spoken dialog [1]–[3]. A critical enabler for effective interaction in this context is E-TTS synthesis. For example, emphasis should be given to important points in the synthesized speech, while different intonations should be applied to different dialog states (e.g., question versus confirmation).

There exists a rich repository of previous work in E-TTS [4]–[7]. Earlier efforts established that recognizable vocal effects can be generated in rule-based synthesis of vocal emotions [8]–[10]. A comprehensive review of vocal emotions and their communication process was presented in [11]. This work also pointed out the lack of a consensual definition of different vocal emotions (e.g., happy, sad, surprise, etc.) and different qualitative types of emotions. Many studies have adopted the categorical definitions of the "big six" emotions (i.e., happy, sad, surprise, fear, angry, and disgust) [5], [12]. The scope of emotions was further extended to *expressions* in [13], which include paralinguistic events. Studies were devoted to the realization of expressions through speech prosody and their acoustic correlates, including intonation, amplitude, duration, timing, and voice quality [14]–[17]. Large databases of expressive speech have also been collected to support data-driven research [18], including the Reading–Leeds Emotion Speech project [19], the Belfast project [20], and the CREST-ESP project [21]. Explorations have been undertaken in the use of concatenative methods for E-TTS [22], [23]. Speech recordings from different emotion categories were utilized with TD-PSOLA to mix and match the prosodic information and diphone inventories for different emotion states [24], [25]. Results show that consistent selection of the prosodic and diphone inventories according to the intended emotion for synthesis gives the highest emotion accuracies. Another engineering approach converts prosody-related acoustic features from neutral to emotional speech, using methods such as the linear modification model (LMM), Gaussian mixture model (GMM), and classification and regression trees (CART) [26]. Additionally, the work in [14] presented a continuum of emotion states in synthetic speech using psychological emotion dimensions (i.e., activation, evaluation, and power). The work also demonstrated the possibility of synthesizing emotional speech acoustics that correspond to different locations in the three-dimensional emotion space. Furthermore, enhancement of expressivity has also

Manuscript received September 03, 2008; revised April 20, 2009. First published May 15, 2009; current version published August 14, 2009. This work was supported in part by the joint research fund of the National Natural Science Foundation of China—Hong Kong SAR Government Research Grants Council (NSFC-RGC) under Grants 60418012 and N-CUHK417/04. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Abeer Alwan.

Z. Y. Wu and H. M. Meng are with the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong (CUHK), China, and also with the Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China (e-mail: zywu@se.cuhk.edu.hk; hmmeng@se.cuhk.edu.hk).

H. W. Yang is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: yang-hw03@mails.tsinghua.edu.cn).

L. H. Cai is with the Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China, and also with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: clh-dcs@tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2023161

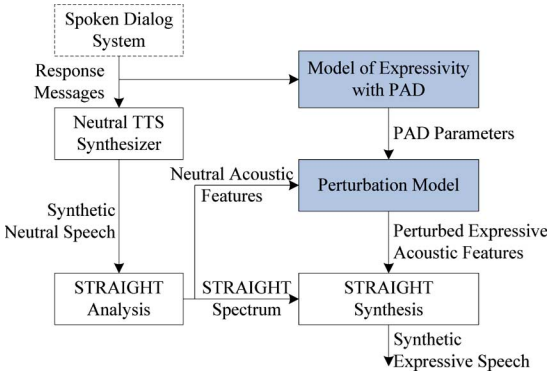

Fig. 1. Overview of the expressive text-to-speech (E-TTS) synthesizer for response generation in a spoken dialog system.

been attempted through audiovisual means [27]—by utilizing the relative timing and influence among audiovisual cues.

This paper seeks to incorporate expressivity in synthetic speech to convey the *communicative semantics* of system response messages in a spoken dialog system. This problem may be divided into two parts: 1) to develop a mapping from the text semantics in the system response messages to a set of descriptors for expressivity; as well as 2) to develop a perturbation model that can render acoustic features of speech according to the parameterized descriptors of expressivity. In relation to the first subproblem, the conventional categories of vocal emotion cannot be applied in a straightforward manner. Instead, we adapt Mehrabian's three-dimensional PAD model [28] as the *descriptive framework* for semantic-oriented expressions. The three dimensions are approximately orthogonal and the abbreviations include: P for pleasure-displeasure, A for arousal-nonarousal, and D for dominance-submissiveness. The PAD model has been successfully applied to describe emotions, moods, situations [30], and personalities [31]. In relation to the second subproblem, we present a *non-linear perturbation model* that can postprocess the neutral TTS output obtained from the existing spoken dialog system to generate expressive TTS outputs for enhanced interactions. Fig. 1 presents an overview of the E-TTS synthesizer.

The rest of the paper is organized as follows. We first describe the scope of our investigation which is defined in the tourist information domain, followed by an introduction of the PAD framework and its parameterization for describing the expressivity of text semantics. Then we present the nonlinear perturbation model for rendering expressive acoustics, followed by experiments, analysis of results and conclusions.

II. Scope

This research is conducted in the context of a spoken dialog system in the tourist information domain [32]. Information is sourced from the Discover Hong Kong website of the Hong Kong Tourism Board [33]. The use of E-TTS to generate system responses aims to enhance interactivity between human and computer. We collected dialog data from 30 recruited subjects with a Wizard-of-Oz (WoZ) setup. Each subject interacts with a browser-based interface, behind which "hides" the wizard who can access the Discover Hong Kong website freely. The user's inquiries may be presented with speech, typed text and mouse gestures. The wizard attempts to provide cooperative and informative responses in terms of speech and tagged information (e.g., URLs, highlighted locations on a map, etc.) All dialog interactions were logged by the system. Analysis shows that the wizard's speech may be spontaneous and contain disfluencies. To ease subsequent processing, we devised a procedure of data regularization which simplifies the wizard's responses into utterances with straightforward grammar structures [32]. Overall we collected 1500 dialog turns, each of which contains two to five utterances. This amounts to 3874 utterances in user inquiries and system responses. Tables I and II illustrate that the tourist information domain presents four genres of response messages: [1]

1) the *interactive* genre that characterizes dialog interactions (e.g., carry forward to the next dialog turn or bring it to a close);
2) the *descriptive* genre that describes the attractive features of a scenic spot;
3) the *informative* genre that presents facts (e.g., opening hours and/or ticket price of a scenic spot);
4) the *procedural* genre that gives instructions (e.g., transportation and working directions).

Utterances in the interactive genre (see Table I) have been semi-automatically annotated with key concepts (KC), dialog acts (DA), and task goals (TG). KCs correspond to the lexical semantics of Chinese words and are extracted by homegrown tokenization and parsing algorithms. The DA denotes the communicative goal of an utterance in the context of the dialog and bears relationships with neighboring dialog turns. We use DAs that are adapted from VERBMOBIL-2 [34] . The TG refers to the user's informational goal that underlies an utterance and our

TABLE I
Example Dialog Between User (U) and System (S). Key Concepts (KC) and Dialog Acts (DA) Have Been Labeled Semi-Automatically. Utterances Belong to the Interactive Genre

S0	請問，你第二天想去哪裡？ *(Where would you like to go on the second day?)* KC: {ask_where=去哪裡 *(go to where)*} TG: ATTRACTION DA: REQUEST_PREFERENCE
U1	我想去主題公園看看。*(I would like to visit a theme park.)* KC: {attraction=主題公園 *(theme park)*} TG: ATTRACTION DA: INFORM_DETAILS
S1	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?)* KC: {attraction=海洋公園 *(Ocean Park)*, attraction=迪士尼樂園 *(Disneyland)*} TG: ATTRACTION DA: REQUEST_COMMENT
U2	讓我想想。*(Let me think.)* KC: {think=想想 *(think)*} TG: ATTRACTION DA: DEFER
U2	去海洋公園好了。*(Ocean Park.)* KC: {attraction=海洋公園*(Ocean Park)*} TG: ATTRACTION DA: INFORM_DETAILS
S2	這裡是海洋公園的資料，請看。 *(Here is some information about Ocean Park)* KC: {attraction=海洋公園*(Ocean Park)*} TG: ATTRACTION DA: INFORM_GENERAL 香港海洋公園是全東南亞規模最大的水族及主題公園之一…… *(May be followed by more information about Ocean Park from the Discover Hong Kong website.)*

[1]English translations are provided in the tables for readability.

TABLE II
TYPICAL INFORMATION GIVEN ABOUT A SCENIC SPOT. (SOURCE: DISCOVER HONG KONG WEBSITE [33])

<name of tourist spot>海洋公園 *(Ocean Park)*
<descriptive genre>
香港海洋公園是全東南亞規模最大的水族及主題公園之一。您可以在園內參觀大型的珊瑚礁水族館，觀賞海豚表演，探訪可愛的國寶大熊貓“安安”和“佳佳”，嘗試機動城內多種緊張刺激的機動遊戲，玩個不亦樂乎。
(One of Southeast Asia's largest oceanariums and theme parks, featuring aquariums, dolphin shows, thrilling rides, giant pandas An An and Jia Jia, and much more.)
<informative genre>
開放時間為每日上午十點至下午六點。
(Open daily, 10am-6pm.)
<procedural genre>
您可於地鐵金鐘站B出口或中環天星碼頭附近（地鐵中環站K出口）乘城巴629路往返海洋公園
(Special City bus 629 leaves from near the Star Ferry Piers (Central MTR Exit K) in Central and Admiralty MTR Exit B.)

TGs are designed specifically for the tourist information domain. We use trained belief networks to infer the DA and TG in the dialog corpus [32]. Speech synthesis for a response message in the interactive genre will need to incorporate appropriate utterance-level intonation for different DA.

Utterances in the three remaining genres (see Table II) aim to provide information for the user. The descriptive genre often contains commendatory words that describe scenic spots and their specialties. The informative and procedural genres contain useful facts for the tourist. Speech synthesis for a response message in these genres will need to incorporate appropriate word-level prosody based on lexical semantics, with suitable emphasis to draw the attention of the listener.

III. TEXT PROMPTS AND SPEECH CORPUS

We collected contrastive (i.e., neutral versus expressive) speech recordings of text prompts that cover the four different genres of response messages, namely, the interactive, descriptive, informative, and procedural genres.

A. Text Prompts

Text prompts that belong to the interactive genre include 1063 response messages selected from the WoZ dialogs (see previous section). These text prompts consist of 6047 Chinese prosodic words and 13 555 Chinese syllables. Text prompts that belong to the descriptive, informative, and procedural genres are derived from text passages corresponding to 20 scenic spots in the Discover Hong Kong website. The text passages include 60 paragraphs, consisting of 357 utterances, 1358 Chinese prosodic words and 3340 Chinese syllables. The prosodic word is defined as the smallest constituent at the lowest level of the prosodic hierarchy, and consists of a group of syllables uttered closely and continuously in an utterance [35], [36]. We have chosen the prosodic word as the basic unit for analysis and modeling since it provides a natural connection between the text semantics and speech acoustics.

B. Speech Corpus

A male native Mandarin speaker was recruited to record in a soundproof studio. The speaker has several years of research experience in expressive speech processing. Therefore, he has considerable understanding of the differences between neutral and expressive speech. For each text prompt, the speaker was asked to record contrastive versions of neutral and expressive speech. For text prompts that belong to the descriptive, informative, and procedural genres, expressive speech recordings should contain local, word-level expressivity that conveys the lexical semantics of the prosodic words. For the text prompts that belong to the interactive genre, expressive speech recordings should contain global expressivity that conveys the communicative goal (i.e., dialog act) of the utterance. We have 60 text prompts that fall under the descriptive, informative, and procedural genres. These prompts tend to be long and each may contain one to eight sentences, leading to 357 utterances in total. We have 1063 text prompts in the interactive genre and each corresponds to one utterance. Altogether the recordings amount to 225 min of speech. The sound files are saved in the .wav format (16 bit mono, sampled at 16 kHz). This data is needed for data analysis and modeling. We set aside another disjoint set of 60 utterances from the descriptive, informative, and procedural genres and 60 from the interactive genre, to be used as the test set for experimentation.

IV. MODELING EXPRESSIVITY WITH THE PAD FRAMEWORK

As mentioned in Section I, the first of our two subproblems is to develop a mapping from text semantics in response messages to a set of descriptors for expressivity. We find that conventional emotion categories do not offer a sufficiently general descriptive framework for semantic-oriented expressivity. Instead, we adapt Mehrabian's PAD model [28] which has three approximately orthogonal dimensions: 1) "pleasure-displeasure" (P) distinguishes the positive-negative affective qualities of emotion states; 2) "arousal-nonarousal" (A) refers to a combination of physical activity and mental alertness; and 3) "dominance-submissiveness" (D) is defined in terms of control versus lack of control. The axis for each dimension ranges from -1.0 to 1.0. It has been shown in [29] and [30] that the PAD space provides an adequate and general characterization of emotions, covering 240 emotion states. Previous work in psychology has also devised elicitation methods to obtain PAD values for emotion terms [30], e.g., "elated" corresponds to $(P = 0.50,\ A = 0.42,\ D = 0.23)$ and "inhibited" to $(P = -0.54,\ A = -0.04,\ D = -0.41)$. PAD values can also be used to describe situations [28]. For example, the situation "you have had a long and exhausting day at work; you now must wait for about 30 to 40 min for your ride home" corresponds to all negative PAD values (i.e., –P–A –D). For the current investigation, we believe that the PAD model offers a general description framework that can cover local expressivity at the word level based on lexical semantics, as well as global expressivity at the utterance level based on dialog acts.

A. Heuristics for PAD Parameterization

We designed a set of heuristics such that the PAD descriptors can be parameterized according to the semantic content of the response messages. The heuristics are applied at *two levels*: the P and A descriptors are used for local expressivity at the prosodic word level based on lexical semantics, while the D descriptor is used for global expressivity at the utterance level

TABLE III
CORRESPONDENCES BETWEEN DIALOG ACTS (DA) AND THEIR D VALUES

Dialog Act (DA)	Example utterances	D value
CONFIRM FEEDBACK_POSITIVE FEEDBACK_NEGATIVE	你的票已經訂好了。 *(Your ticket has been booked)*	1
CLARIFY CLOSE BYE INFORM_DETAILS INFORM_GENERAL	這裡就是海洋公園的資料。 *(Here is some information about Ocean Park.)*	0.5
BACKCHANNEL OOD (out of domain)	啊 *(hmm…)*	0
COMMIT DEFER THANK SUGGEST	請等一等…… *(Please wait a minute.)*	–0.5
APOLOGY REQUEST_COMMENT REQUEST_PREFERENCE REQUEST_DETAILS REQUEST_CLARIFY REQUEST_ACTION	對不起，我沒有找到有關資料。 *(Sorry, I cannot find the requested information.)* 你想去哪裡呢？ *(Which place do you wish to visit?)*	–1

TABLE IV
CORPUS STATISTICS OF THE ANNOTATED D VALUES BASED ON THE DIALOG ACTS (DA) OF UTTERANCES IN THE INTERACTIVE GENRE

D	–1	–0.5	0	0.5	1
# of utterances (total: 1,063)	351	145	141	316	110
% of occurrences	33.0	13.6	13.3	29.7	10.4

TABLE V
CORPUS STATISTICS OF THE ANNOTATED P AND A VALUES BASED ON THE LEXICAL SEMANTICS OF PROSODIC WORDS IN THE INTERACTIVE GENRE

(P, A)	(0, 0)	(0, 0.5)	(–1, 0.5)	(1, 0.5)	(1, 1)
# of prosodic words (total: 6,047)	3,161	2,419	66	339	62
% of occurrences	52.3	40.0	1.1	5.6	1.0

TABLE VI
CORPUS STATISTICS OF THE ANNOTATED P AND A VALUES BASED ON THE LEXICAL SEMANTICS OF PROSODIC WORDS IN THE DESCRIPTIVE/INFORMATIVE/PROCEDURAL GENRES

(P, A)	(0, 0)	(0, 0.5)	(1, 0.5)	(1, 1)
# of prosodic words (total: 1,358)	238	922	141	57
% of occurrences	17.5	67.9	10.4	4.2

TABLE VII
EXAMPLE OF P AND A ANNOTATIONS FOR PROSODIC WORDS IN AN UTTERANCE

prosodic words	這個 *(this)*	呈半月型的 *(crescent-shaped)*	沙灘 *(beach)*	是香港 *(is Hong Kong's)*	最受 *(most)*	歡迎的 *(popular)*	海灘之一 *(beach)*
(P, A)	(0, 0)	(0, 0.5)	(0, 0.5)	(0, 0.5)	(1, 1)	(1, 1)	(0, 0.5)

based on dialog acts. We elaborate on the details of the heuristic principles and their motivations as follows.

1) *P Values* Commendatory words or words with positive connotations are labeled with $P = 1$, e.g., 歡迎的 (*popular*), 美麗 (*beautiful*), etc. Derogatory words or words with negative connotations are labeled with $P = -1$, e.g., 偏僻 (*devious*), 擁擠 (*crowded*), etc. Remaining words are assumed neutral and are labeled with $P = 0$.
2) *A Values*: Superlatives and words denoting a high degree are labeled with the maximum level of arousal as $A = 1$, e.g., 最 (*most*), 非常 (*very*), 極 (*super*), etc. Comparatives and words carrying key facts are labeled with an intermediate level of arousal as $A = 0.5$, e.g., 比較 (*relative*), street names, transportation means, etc. Hence, these words carry a moderate amount of emphasis. The remaining words are labeled with $A = 0$. A common sentence construct found in the text prompts 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) *D values*: Utterances that provide confirmation or feedback are labeled with $D = 1$ (i.e., very dominant). Utterances that give introductions, explain facts or bring dialogs to a close are labeled with $D = 0.5$ (i.e., moderately dominant). Utterances that give suggestions, express thanks, ask for help or deferment are labeled with $D = -0.5$ (i.e., submissive). Apologetic utterances and interrogative utterances are labeled with $D = -1$ (i.e., very submissive). The remaining utterances are labeled with $D = 0$. Table III presents some examples of the correspondences between dialog acts (DA) and their D values.

Recall from Fig. 1 that E-TTS is applied to the response message text that is generated by a spoken dialog system [32]. Message generation is performed with a template-based approach. The heuristics presented above can be easily incorporated into the templates for PAD parameterization based on response message text.

B. Corpus Statistics Based on PAD Annotations

Three annotators were asked to follow the heuristic principles to annotate the response message texts. An agreement for 94% of the prosodic words is achieved among the three annotators in the annotation of P and A values. Ambiguity is resolved by majority rule or a further pass in annotation. Annotation of D value of an utterance is a straightforward mapping based on the dialog act inferred by trained belief networks. Hence, there is no ambiguity. Corpus statistics based on the annotations are shown in Tables IV–VI. The tourist information domain contains primarily pleasant prosodic words about scenic spots. Table VII gives an example of annotated P and A values for prosodic words in an utterance. Due to the sparseness of prosodic words with negative P value, they are excluded from the subsequent study.

V. EXPLORATORY DATA ANALYSIS OF THE ACOUSTIC CORRELATES OF EXPRESSIVITY

Having adopted the PAD framework to produce a mapping from text semantics of the dialog response messages to the parameterized descriptors for expressivity; we proceed with an exploratory data analysis of the acoustic correlates of the descriptors. We capture both the average and the dynamicity of acoustic features commonly associated with expressive speech:

- **Intonation**: f_0 mean, f_0 range, f_0 slope;

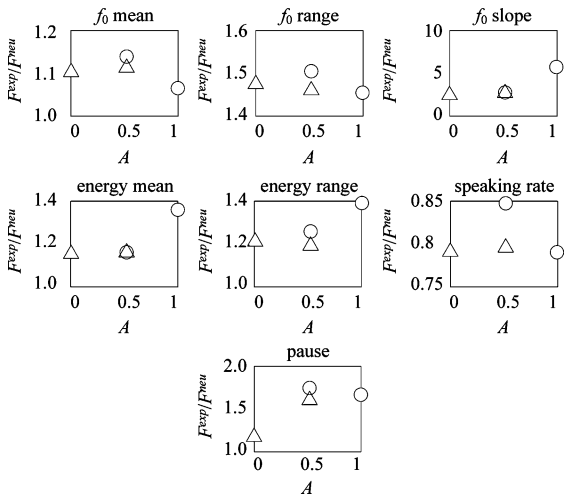

Fig. 2. Ratio between expressive and neutral speech acoustic features $F^{\mathrm{exp}}/F^{\mathrm{neu}}$ for different P and A values. Triangles denote $(P = 0)$ and circles denote $(P = 1)$.

- **Intensity**: mean and range of the root mean square (RMS) energy (energy mean and energy range);
- **Speaking rate**: syllables per minute; and
- **Fluency**: pause duration before a prosodic word.

The analysis is conducted based on the contrastive recordings from our speech corpus. Each recorded utterance is automatically segmented into syllables by forced alignment with an HMM-based speech recognizer and the syllable boundaries are checked manually. Measurements are then taken from each syllable. We compute the ratio of each feature value $(F^{\mathrm{exp}}/F^{\mathrm{neu}})$ between the neutral (F^{neu}) and expressive counterparts (F^{exp}), where F denotes any of the features described above.

To understand local expressivity, we analyze recordings from the descriptive, informative, and procedural genres. The prosodic variations in these utterances are primarily due to local, word-level expressivity for conveying lexical semantics. There is relatively little variations due to utterance-level dialog acts (since most of the utterances carry $D = 0.5$). Fig. 2 depicts the seven acoustic features for difference combinations of (P, A) values. Here, the A values are shown on the x-axis, triangles denote cases when $(P = 0)$ and circles denote cases when $(P = 1)$. We observe that all ratios between expressive and neutral speech acoustic features, except for speaking rate, are larger than one (i.e., $F^{\mathrm{exp}}/F^{\mathrm{neu}} > 1$). This agrees with common perception that expressive speech has higher values for f_0 mean and energy, and lower values for speaking rate. We also observe that when $P = 1$ (referring to the circles in the figure), the f_0 range and speaking rate decrease as the A value increases. This also agrees with common perception that speakers may emphasize certain words by speaking more slowly with a steady intonation.

To understand global expressivity, we analyze recordings from the interactive genre. These utterances should carry low prosodic variations due to word-level expressivity because the majority (over 92%) of the words have neutral (P, A) values (i.e., $P = 0$ and $A = 0$ or 0.5, see Table V). Instead, the range of D values covered by this dataset implies that prosodic variations should primarily be due to utterance-level dialog acts.

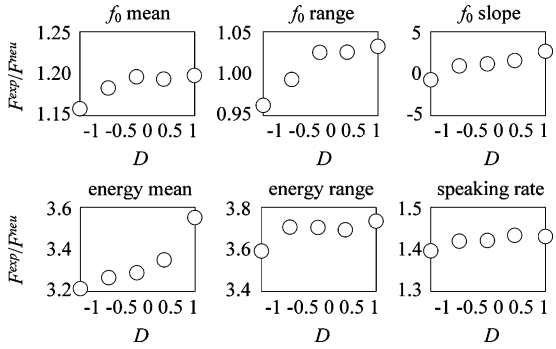

Fig. 3. Ratio between expressive and neutral speech acoustic features $F^{\mathrm{exp}}/F^{\mathrm{neu}}$ for different D values.

Fig. 3 shows variations in six acoustic features. Pause durations in between prosodic words are ignored because they associate mainly with the local level. We observe that the ratio between expressive and neutral speech acoustic features increases with the D value. This agrees with the common perception that dominance often leads to exaggerated expressions. In particular, f_0 has a negative slope at $D = -1$, corresponding to rising intonation (i.e., interrogative intonation) at the utterance-final position.

VI. Perturbation Model for Expressive Synthesis

We proceed to develop a model that can render expressive speech acoustic features based on parameterized descriptors of expressivity. Based on the exploratory data analysis, we observe a nonlinear relationship between PAD descriptors and their acoustic correlates. Hence, we propose a nonlinear perturbation model for transforming neutral speech acoustic features into expressive renditions. The approach involves *two levels* of perturbation: 1) the prosodic word level based on the lexical semantics and their (P, A) values; and 2) the utterance level based on the dialog acts and their D values.

A. Local Perturbation at the Prosodic Word Level

The model for local perturbation is driven by (P, A) values, as shown as follows:

$$\frac{F^{\mathrm{exp}}}{F^{\mathrm{neu}}} = C_1 P \exp(-C_2 A) + C_3 A \exp(-C_4 P) + C_5 \quad (1)$$

where F^{exp} denotes any of the seven features (see Section V) from expressive speech, F^{neu} is the corresponding feature from neutral speech, $F^{\mathrm{exp}}/F^{\mathrm{neu}}$ is the ratio between expressive and neutral speech acoustic feature, and $C_1, \ldots, C_5$ are coefficients. This equation captures the observed trends when $P = 0$, the expressive measurement increases linearly with A (from 0 to 0.5). However, the linear relationship changes to exponential when $P = 1$. This is captured by the factor $C_1 \exp(-C_2 A)$ for increments of A from 0.5 to 1. Nonlinear least-squares regression[2] is used to estimate the coefficients from utterances in the descriptive, informative, and procedural genres.

[2]Coefficients are initialized at 1 and maximum number of iterations is set at 100.

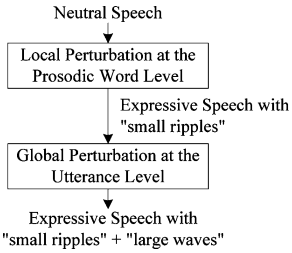

Fig. 4. Integrated local and global perturbation model for expressive speech synthesis.

B. Global Perturbation at the Utterance Level

A similar nonlinear model is used for global perturbation and driven by D values, as shown as follows:

$$\frac{F^{\exp}}{F^{\text{neu}}} = C_6 D \exp(-C_7 D) + C_8 \tag{2}$$

where $F^{\exp}$, F^{neu}, and $F^{\exp}/F^{\text{neu}}$ have the same meaning as in (1), and C_6, C_7, C_8 are coefficients. Similar to the above, nonlinear least-squares regression is used to estimate the coefficients from utterances in the interactive genre.

C. Integrated Local and Global Perturbation

To integrate the local and global perturbations, we refer to Chao's theory [37] for Chinese speech, which states that the expressive intonation is the combination of "small ripples" (syllabic tone) and "large waves" (intonation). Such integration enables us to render expressivity based on different combinations of PAD values, encountered in the different genres of response messages. For example, the dialog response "OK" may confirm the user's statement; while "OK?" seeks confirmation from the user. These two messages with identical textual content will have the same local expressivity due to the prosodic word, but different global expressivity due to the dialog act. "OK" should have declarative intonation while "OK?" should have interrogative intonation.

Fig. 4 illustrates the sequential framework for integration. The first step uses the local perturbation model in (1) to modulate the seven acoustic measurements according to (P, A) values in each prosodic word. Modulation is realized for each syllable by means of STRAIGHT [38]. Pause segments with appropriate durations are concatenated to the beginning of each prosodic word. Modulated, expressive speech segments from all the prosodic words are then concatenated to generate a synthetic speech utterance with local expressivity. The second step further applies global perturbation in (2) to modulate the six acoustic measurements according to the D value of each utterance. Modulation is also realized by STRAIGHT [38].

D. Implementing Perturbations With STRAIGHT

The STRAIGHT algorithm [38] provides an "analysis-modification-resynthesis" framework for converting input speech to target speech with desired characteristics (e.g., new acoustic features or new spectrum). This presents a very desirable platform for the incorporation of the proposed perturbation model. Neutral speech input is first analyzed by STRAIGHT to obtain the spectrum, pitch and energy features, and durations. These acoustic features are then modulated with the proposed perturbation model to generate the target acoustic features for expressive speech. Fig. 1 shows that the final step involves feeding the perturbed acoustic features and spectrum into STRAIGHT to resynthesize expressive speech.

1) Analysis: To describe the process of analysis, we denote the neutral, synthetic speech generated by our existing TTS synthesizer with $S(t)$. Since we use syllable concatenative synthesis, $S(t)$ can also be represented as N syllable waveforms $S_i(t),\ i = 1, \ldots, N$

$$S(t) = \{S_1(t), \ldots, S_i(t), \ldots, S_N(t)\} \tag{3}$$

where t is the discrete time index measured in milliseconds. We denote the known boundaries of ith syllable waveform $S_i(t)$ with $[b_i,\ e_i]$, i.e., begin/end times in milliseconds. We compute the RMS energy for every millisecond of the syllable waveform $S_i(t)$ to generate $E_i(t)$.

STRAIGHT analysis computes the speech spectrum with a 1024-point FFT, with the analysis rate (i.e., window advancement) of 1 ms. From each spectrum, the pitch (f_0) is extracted between the search range of 40 to 800 Hz. We denote the spectrum and pitch contour for syllable $S_i(t)$ with $W_i(t)$ and $P_i(t)$, where $t \in [b_i, e_i]$.

From these measurements, we obtain the seven acoustic features mentioned in Section V. Acoustic features for intonation, i.e., f_0 mean (P_i^{Mean}) and f_0 range (P_i^{Range}) for syllable i are calculated as follows:

$$P_i^{\text{Mean}} = \frac{1}{e_i - b_i + 1} \sum_{t=b_i}^{e_i} P_i(t) \tag{4}$$

$$P_i^{\text{Range}} = \max\left[P_i(t)\right] - \min\left[P_i(t)\right],\ t \in [b_i, e_i]. \tag{5}$$

We apply linear regression $f_i(t) = \alpha_i + \beta_i t$ to the pitch contour $P_i(t)$. The slope β_i is the f_0 slope (P_i^{Slope}):

$$P_i^{\text{Slope}} = \frac{\sum_{t=b_i}^{e_i} (t - \bar{t})\left(P_i(t) - P_i^{\text{Mean}}\right)}{\sum_{t=b_i}^{e_i} (t - \bar{t})^2}$$

$$\bar{t} = \frac{1}{2}(b_i + e_i). \tag{6}$$

The intercept α_i is calculated from f_0 mean and f_0 slope as

$$\alpha_i = P_i^{\text{Mean}} - P_i^{\text{Slope}} \bar{t},\ \bar{t} = \frac{1}{2}(b_i + e_i). \tag{7}$$

Acoustic features for intensity, i.e., energy mean (E_i^{Mean}) and energy range (E_i^{Range}) for syllable i are computed as

$$E_i^{\text{Mean}} = \frac{1}{e_i - b_i + 1} \sum_{t=b_i}^{e_i} E_i(t) \tag{8}$$

$$E_i^{\text{Range}} = \max\left[E_i(t)\right] - \min\left[E_i(t)\right],\ t \in [b_i, e_i]. \tag{9}$$

The duration D_i of syllable i and the duration of its preceding pause Z_i are measured as follows:

$$D_i = e_i - b_i + 1 \tag{10}$$

$$Z_i = b_i - e_{i-1} - 1. \tag{11}$$

2) Modification and Resynthesis: Perturbations at the local and global levels are realized by multiplication with the ratios $F^{\text{exp}}/F^{\text{neu}}$ from (1) and (2), respectively. We denote the expressive rendition for syllable i with $\tilde{S}_i(t)$. Resynthesizing using STRAIGHT requires three parameters: 1) the spectrum of the expressive speech $\tilde{W}_i(t)$, 2) the pitch contour $P_i(t)$, and 3) the time-axis mapping information $T_i(t)$ (we will elaborate on this later). The spectrum and pitch contour are used in STRAIGHT to resynthesize speech in the frequency domain; while the time-axis mapping information is used for changing the speaking rate in the time domain. The spectrum and pitch contour should have the same temporal length.

The spectrum is not modified in our current work. Hence,

$$\tilde{W}_i(t) = W_i(t), t \in [b_i, e_i]. \tag{12}$$

The pitch contour $\tilde{P}_i(t)$ of expressive syllable should have the same temporal length as the spectrum, and hence the same syllable boundaries $[b_i,\ e_i]$ as neutral speech. The new pitch contour $\tilde{P}_i(t)$ is calculated from the neutral pitch contour $P_i(t)$ with two steps. First, the slope of the pitch contour is changed by subtracting the fitted straight line of the neutral pitch contour $f_i(t)$ [see (6) and (7)] followed by incorporating the desired f_0 slope $(\tilde{P}_i^{\text{Slope}})$ as $\hat{f}_i(t)$ with zero as f_0 mean

$$\begin{aligned}
f_i(t) &= \alpha_i + \beta_i t = P_i^{\text{Mean}} - P_i^{\text{Slope}}\bar{t} + P_i^{\text{Slope}} t \\
\hat{f}_i(t) &= \hat{\alpha}_i + \hat{\beta}_i t = 0 - \tilde{P}_i^{\text{Slope}}\bar{t} + \tilde{P}_i^{\text{Slope}} t \\
\hat{P}_i(t) &= [P_i(t) - f_i(t)] + \hat{f}_i(t) \\
&= \left[P_i(t) - P_i^{\text{Mean}} - P_i^{\text{Slope}}(t - \bar{t})\right] \\
&\quad + \tilde{P}_i^{\text{Slope}}(t - \bar{t}) \\
&= P_i(t) + \left(\tilde{P}_i^{\text{Slope}} - P_i^{\text{Slope}}\right)(t - \bar{t}) \\
&\quad - P_i^{\text{Mean}} \\
t &\in [b_i, e_i],\ \bar{t} = \frac{1}{2}(b_i + e_i).
\end{aligned} \tag{13}$$

Thereafter, the pitch contour is shifted and scaled to match the desired f_0 mean $(\tilde{P}_i^{\text{Mean}})$ and f_0 range $(\tilde{P}_i^{\text{Range}})$

$$\begin{aligned}
\hat{P}_i^{\text{Range}} &= \max\left[\hat{P}_i(t)\right] - \min\left[\hat{P}_i(t)\right] \\
\tilde{P}_i(t) &= \hat{P}_i(t)\frac{\tilde{P}_i^{\text{Range}}}{\hat{P}_i^{\text{Range}}} + \tilde{P}_i^{\text{Mean}},\ t \in [b_i, e_i].
\end{aligned} \tag{14}$$

We denote the target boundaries for expressive speech of syllable i with $[\tilde{b}_i,\ \tilde{e}_i]$, which are computed as

$$\begin{aligned}
\tilde{b}_i &= \sum_{j=1}^{i-1}\tilde{D}_j + \sum_{j=1}^{i}\tilde{Z}_j \\
\tilde{e}_i &= \tilde{b}_i + \tilde{D}_i.
\end{aligned} \tag{15}$$

The durations $\tilde{D}_j$ and $\tilde{Z}_j$ can be computed from the perturbation models, i.e., (1) and (2). Then we can obtain the time-axis mapping information $T_i(t)$ which maps the time index on expressive speech to the time index on neutral speech, to find appropriate parameters for resynthesizing expressive speech:

$$T_i(t) = \frac{e_i - b_i}{\tilde{e}_i - \tilde{b}_i}\left(t - \tilde{b}_i\right) + b_i,\ t \in \left[\tilde{b}_i, \tilde{e}_i\right]. \tag{16}$$

STRAIGHT begins with resynthesis of $\hat{S}_i(t)$ (without energy modification) based on $\tilde{W}_i(t)$, $\tilde{P}_i(t)$, and $T_i(t)$:

$$\hat{S}_i(t) = f\left(\tilde{W}_i(t), \tilde{P}_i(t), T_i(t)\right),\ t \in \left[\tilde{b}_i, \tilde{e}_i\right] \tag{17}$$

where $f(\cdot)$ represents the synthesis process of the STRAIGHT algorithm, details of which are presented in [38].

Thereafter, energy level of $\hat{S}_i(t)$ is adjusted by scaling with the desired energy mean $(\tilde{E}_i^{\text{Mean}})$ and range $(\tilde{E}_i^{\text{Range}})$:

$$\tilde{E}_i(t) = \left[E_i(t) - E_i^{\text{Mean}}\right]\frac{\tilde{E}_i^{\text{Range}}}{E_i^{\text{Range}}} + \tilde{E}_i^{\text{Mean}},\ t \in \left[\tilde{b}_i, \tilde{e}_i\right]. \tag{18}$$

Then, the energy of $\hat{S}_i(t)$ is scaled by $\tilde{E}_i(t)/E_i(t)$ and further smoothed by a Hamming window $\omega_i(t)$ in preparation for syllable waveform segment concatenation to produce the final expressive speech segment $\tilde{S}_i(t)$ of syllable i:

$$\tilde{S}_i(t) = \hat{S}_i(t)\frac{\tilde{E}_i(t)}{E_i(t)}\omega_i(t),\ t \in \left[\tilde{b}_i, \tilde{e}_i\right]. \tag{19}$$

The Hamming window $\omega_i(t)$ is defined as

$$\omega_i(t) = 0.53836 - 0.46164\cos\left(\frac{2\pi(t - \tilde{b}_i)}{\tilde{e}_i - \tilde{b}_i}\right),\ t \in [\tilde{b}_i, \tilde{e}_i]. \tag{20}$$

Finally, the entire expressive speech is generated by concatenating the N syllable waveforms:

$$\tilde{S}(t) = \{\tilde{S}_1(t), \ldots, \tilde{S}_i(t), \ldots, \tilde{S}_N(t)\}. \tag{21}$$

VII. Perceptual Evaluation

We conducted a set of perceptual experiments to evaluate the expressive speech synthesized by the integrated perturbation framework. To minimize learning effects which may affect the evaluation results, we divided the test set (as described in Section III-B) into three non-overlapping subsets and conducted three evaluations with one-month intervals. The first subset contains 20 utterances from the descriptive, informative, and procedural genres, which aims to focus on perceiving expressivity at the prosodic word level. The second contains 20 utterances from the descriptive, informative, and procedural genres, as well as 30 utterances from the interactive genre, which focuses on perceiving expressivity at the utterance level. All remaining testing data are grouped into the third subset. Preprocessing of the text prompts includes applying a homegrown tool for prosodic word tokenization, trained belief networks for dialog act inference, as well as the heuristic mapping to obtain the PAD values for prosodic words and utterances. We also verified that all the data subsets have good coverage of the possible combinations in the

TABLE VIII
PERCEPTUAL EVALUATION OF LOCAL PERTURBATION, MEASURED BY THE % OF PROSODIC WORDS JUDGED TO BE CLOSER TO THE EXPRESSIVE RATHER THAN NEUTRAL RECORDINGS

(P, A)	(0, 0)	(0, 0.5)	(1, 0.5)	(1, 1)
# of prosodic words	17	76	14	15
% of prosodic words	70.6	73.2	84.5	76.1

PAD space. Thereafter, E-TTS is applied to generate expressive utterances from the text prompts.

A. Local Perturbation on Neutral Speech Recordings

We use the first testing data subset to focus on local expressivity for lexical semantics in each prosodic word. We recruited 14 native speakers of Mandarin (nine male, five female, without hearing impairment) to be subjects for the listening test. All the subjects are engineering students who have experiences with speech and language technologies but not in expressive speech synthesis. Each text prompt was presented to the subject in the form of three speech files:

1) the neutral speech recording from the original male speaker who recorded the speech corpus (see Section III);
2) the expressive speech recording from the same speaker;
3) a locally perturbed signal originating from the neutral speech recording 1).

The three speech files were played for the subjects in the order of 1)-2)-3)-1)-2)-3). The subject was presented with the prosodic words of the text prompt while listening, and was asked to judge whether a prosodic word in 3) sounded more similar to its counterpart 1) or 2). Results shown in Table VIII indicate that over 76% of the locally perturbed prosodic words are perceived to be closer to their expressive counterparts than the neutral one. This reflects that the local perturbation model can effectively synthesize expressivity for lexical semantics.

B. Integrated Perturbation on Neutral Speech Recordings

We conducted another listening test with the second testing data subset to focus on the integrated (i.e., both local and global) perturbation model. Each text prompt was presented to the subject in the form of five speech files:

1) the neutral speech recording from the male speaker who recorded the speech corpus;
2) the expressive speech recording from the same speaker;
3) a locally perturbed speech signal from 1);
4) a globally perturbed speech signal from 1); and
5) an integrated (both locally and globally) perturbed speech signal from 1).

The same 14 subjects were recruited for listening evaluation. The speech files were played either in the order 1)-2)-(x) or 2)-1)-(x). Order selection was randomized. (x) refers to perturbed speech and may either be 3), 4) or 5). The subject was presented with the text prompt while listening, and was asked to judge whether utterance (x) sounded more similar to its counterpart of 1) or 2). Results shown in Table IX indicate that local perturbation generates appropriate expressivity for over 73% of the utterances, global perturbation generates appropriate expressivity for 65% of the utterances, and integrated perturbation offers further enhancements to over 83%.

TABLE IX
PERCEPTUAL EVALUATION OF LOCAL, GLOBAL, AND INTEGRATED PERTURBATIONS, MEASURED BY THE % OF UTTERANCES JUDGED TO BE CLOSER TO THE EXPRESSIVE VERSUS NEUTRAL RECORDING

	Local	Global	Integrated
# of utterances	512	455	587
% of utterances	73.1	65.0	83.9

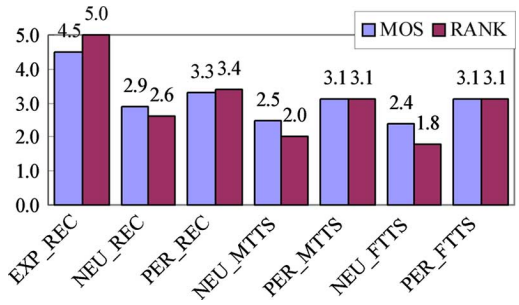

Fig. 5. Comparison between neutral speech recordings, neutral synthetic speech, and their perturbed renditions, based on MOS and absolute ranking.

C. Integrated Perturbation on Neutral TTS Outputs

Thus far, perturbation is applied to neutral recordings provided by the male speaker in our speech corpus. In the spoken dialog system (see Fig. 1), perturbation should be applied to neutral synthetic speech generated by the existing speech synthesizer. This synthesizer is based on the concatenative approach and utilizes voice libraries developed from different speakers (male and female). To assess the extensibility of the proposed perturbation framework, we devised an evaluation that compares between perturbation of neutral speech recordings and neutral synthetic speech. This evaluation involves mean opinion scores (MOS) provided by the same 14 subjects. Each text prompt was presented in the form of seven speech files:

- the neutral and expressive speech recordings from the original male speaker (denoted as NEU_REC and EXP_REC respectively);
- neutral *synthetic* speech in a male or female voice (denoted as NEU_MTTS and NEU_FTTS, respectively);
- the signal obtained from integrated perturbation of NEU_REC, denoted as PER_REC;
- the signal obtained from integrated perturbation of NEU_MTTS, denoted as PER_MTTS; and
- the signal obtained from integrated perturbation of NEU_FTTS, denoted as PER_FTTS.

Subjects were asked to score each speech file based on a five-point Likert scale:

5 Expressive—natural and expressive like human speech;
4 Natural—appropriate for the semantics of the message;
3 Acceptable—flat intonation with some expressivity;
2 Unnatural—robotic with little expressivity;
1 Erratic—low intelligibility and weird.

Results are shown in Fig. 5. Integrated perturbation applied to NEU_REC, NEU_MTTS and NEU_FTTS increases the average MOS by 0.4, 0.6 and 0.7, respectively. These increments

are shown to be statistically significant based on paired t-test with $\alpha = 0.01$. [3] We also observe variations in the range of MOS across subjects. Some subjects never give the full score of 5, or the lowest score of 1. To normalize for such variations across subjects, we also mapped the MOS scores into absolute rankings[4] for the seven speech files. Comparative trends remain consistent, as shown in Fig. 5. These results demonstrate the efficacy and extensibility of the integrated perturbation framework as we migrate from inputs of neutral speech recordings to neutral synthetic speech.

VIII. Discussion

Previous work in [26] attempted to synthesize four types of emotional speech (namely happiness, sadness, fear, and anger) at three levels (i.e., strong, medium, and weak). This amounts to about 12 categories in all. It was found that the performance of the linear modification model (LMM) that maps neutral speech to each emotional category is inferior to the approaches of GMM and CART. The main reason is because the two latter approaches involve finer partitioning of the prosodic space based on stress and linguistic information. The finer partitions help achieve better models for prosodic conversion to synthesize emotional speech. Although the current work focuses on expressive synthesis based on text semantics, our findings seem to be consistent with the previous work. For example, our global perturbation model aims to modulate neutral speech into one of five categories (depending on the D value). We performed a listening test that evaluates the effectiveness of global perturbation in isolation. Results (see Table IX) show that global perturbation generates improved expressivity for 65% of the testing utterances, which is inferior to local perturbation (73%). However, when both perturbations are used in conjunction, a further improvement to 84% is observed. We believe that this improvement is due to the finer partitioning of the prosodic space based on the (P, A) parameters at the lexical word level.

Another noteworthy point is that the current scope of the tourist information domain involves limited variability in PAD values. Hence, the simple heuristic mapping from text semantics to PAD values seems to suffice at the present time, which results in a sparse sampling of PAD combinations. It is conceivable that we can estimate an individual perturbation function for each PAD combination in the current set (four combinations of P and A values, with five D values, totally 20 combinations of PAD values). However, we choose to present a more general framework where the perturbation functions are defined in terms of the P, A and D parameters. This framework is extensible to accommodate higher variability across the PAD continuum as the scope of our domain expands or should we migrate to another (more complex) domain. Under that situation, we will need to adapt psychologically motivated methods for eliciting incremental gradations in the PAD space [28], [30].

[3]The t-test has 13 degrees of freedom since we pair up the corresponding average MOS for each subject.

[4]MOS are ranked in descending order. Tied scores will be assigned the averaged rank, e.g., if speech files B and C have tied scores and should map to ranks 2 and 3, then they will both ranked at 2.5.

Additionally, one may observe that the current perturbation framework achieves an average MOS of about 3, which lies significantly below the desirable upper bound of 5. We believe that further incorporation of fine-grained linguistic information will bring about improvements in performance. As an example, consider the two consecutive prosodic words "最受 (*the most*) " and "歡迎的 (*popular*) "—higher emphasis should be placed on the superlative "最" and the adjective "歡迎" than on the function words (or syllables) "受" and "的". This will be addressed in our future work.

IX. Conclusion and Future Work

This work aims to enhance human–computer interaction in a spoken dialog system by the use of expressive text-to-speech (E-TTS) synthesis to generate system responses. Expressivity in the synthetic speech aims to convey the communicative semantics in the system response text. We organize this research into two parts: 1) to develop a mapping between the text semantics in the response messages to a set of descriptors for expressivity; and 2) to develop a perturbation model that can render acoustic features of expressive speech according to the parameterized descriptors. We propose to adapt the three-dimensional PAD (pleasure-arousal-dominance) model for describing local, word-level expressivity for lexical semantics, as well as global, utterance-level expressivity for dialog acts. We designed a set of heuristics to parameterize the PAD values based on the text semantics of a response message. We also conducted an exploratory data analysis based on contrastive (neutral versus expressive) speech recordings, to understand the acoustic correlates of expressivity at both local and global levels. The analysis led to the development of a nonlinear perturbation model that can transform input neutral speech into expressive speech. Transformation involves local perturbation at the prosodic word level to synthesize expressivity based on lexical semantics; followed by global perturbation at the utterance level to synthesize expressivity based on the dialog act. Perceptual tests using neutral speech recordings show that local perturbation generates appropriate expressivity for 76% of the prosodic words and 73% of the utterances in the test set. Further integration with global perturbation generates appropriate expressivity for 84% of testing utterances. In addition, we compared perturbation of neutral speech recordings with neutral, synthetic speech based on mean opinion scores (MOS). Results show that the integrated perturbation framework improves the average MOS significantly based on paired t-test with $\alpha = 0.01$ for not only neutral speech recordings, but also synthetic speech from different speakers. This presents statistically significant evidence to demonstrate the efficacy and extensibility of the integrated perturbation framework for E-TTS synthesis.

As has been discussed in Section VIII, future investigation will include the incorporation of fine-grained linguistic information (e.g., syntax) in the perturbation framework to achieve performance improvements.

Acknowledgment

The authors would like to thank the anonymous reviewers for their helpful comments.

REFERENCES

[1] R. W. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1997.

[2] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: Facial movements accompanying speech," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognition*, 2002, pp. 381–386.

[3] E. Casatto, J. Ostermann, H. P. Graf, and J. Schroeter, "Lifelike talking faces for interactive service," *Proc. IEEE*, vol. 91, no. 9, pp. 1406–1429, 2003.

[4] N. Campbell, "Towards synthesizing expressive speech: Designing and collecting expressive speech data," in *Proc. Eurospeech*, 2003, pp. 1637–1640.

[5] W. Hamza, E. Eide, R. Bakis, M. Picheny, and J. Pitrelli, "The IBM expressive speech synthesis system," in *Proc. ICSLP*, 2004, pp. 2577–2580.

[6] M. Bulut, S. Narayanan, and L. Johnson, "Synthesizing expressive speech: Overview, challenges, and open questions," in *Text-to-Speech Synth.: New Paradigms Advances*, S. Narayanan and A. Alwan, Eds. Upper Saddle River, NJ: Prentice-Hall, 2004, pp. 175–201.

[7] R. Tsuzuki, H. Zen, K. Tokuda, T. Kitamura, M. Bulut, and S. Narayanan, "Constructing emotional speech synthesizers with limited speech database," in *Proc. ICSLP*, 2004, pp. 1185–1188.

[8] J. E. Cahn, "The generation of affect in synthesized speech," *J. Amer. Voice I/O Soc.*, vol. 8, pp. 1–19, 1990.

[9] I. R. Murray and J. L. Arnott, "Synthesizing emotions in speech: Is it time to get excited?," in *Proc. ICSLP*, Philadelphia, PA, 1996, pp. 1816–1819.

[10] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoust. Soc. Amer.*, vol. 93, no. 2, pp. 1097–1108, 1993.

[11] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.: Special Iss. Speech Emotion*, vol. 40, no. 1–2, pp. 227–256, 2003.

[12] J. C. Martin, S. Abrilian, L. Devillers, M. Lamolle, M. Mancini, and C. Pelachaud, "Levels of representation in the annotation of emotion for the specification of expressivity in ECAs," in *Proc. Intell. Virtual Agents (IVA)*, 2005, pp. 405–417.

[13] G. Bailly, N. Campbell, and B. Möbius, "ISCA special session: Hot topics in speech synthesis," in *Proc. Eurospeech*, 2003, pp. 37–40.

[14] M. Schröder, "Expressing degree of activation in synthetic speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1128–1136, Jul. 2006.

[15] N. Campbell, "Accounting for voice-quality variation," in *Proc. Int. Conf. Speech Prosody*, Nara, Japan, 2004, pp. 217–220.

[16] N. Campbell and P. Mokhtari, "Voice Quality: The 4th prosodic dimension," *Proc. Congr. Phon. Sci.*, pp. 2417–2420, 2003.

[17] T. Banziger and K. R. Scherer, "The role of intonation in emotional expressions," *Speech Commun.*, vol. 46, no. 3–4, pp. 252–267, 2005.

[18] E. Dougles-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Commun.: Special Iss. Speech Emotion*, vol. 40, no. 1–2, pp. 33–60, 2003.

[19] R. Stibbard, "Automated extraction of ToBi annotation data from the Reading/Leeds emotional speech corpus," in *Proc. ISCA Workshop Speech Emotion*, 2000, pp. 60–65.

[20] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A new emotion database: Considerations, sources and scope," in *Proc. ISCA Workshop Speech Emotion*, 2000, pp. 39–44.

[21] N. Campbell, "The JST/CREST ESP project—a midterm progress report," in *Proc. Int. Workshop Expressive Speech Process.*, 2003, pp. 61–70.

[22] A. W. Black, "Unit selection and emotion speech," in *Proc. Eurospeech*, 2003, pp. 1649–1652.

[23] A. Iida, N. Campbell, F. Higuchi, and M. Yasumura, "A Corpus-based speech synthesis system with emotion," *Speech Commun.*, vol. 40, pp. 161–187, 2003.

[24] M. Bulut, S. Narayanan, and A. Syrdal, "Expressive speech synthesis using a concatenative synthesizer," in *Proc. ICSLP*, Denver, CO, 2002.

[25] M. Bulut, C. Busso, S. Yildirim, A. Kazemzadeh, C. M. Lee, S. Lee, and S. Narayanan, "Investigating the role of phoneme-level modifications in emotional speech resynthesis," in *Proc. Eurospeech*, Lisbon, Portugal, 2005.

[26] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1145–1154, Jul. 2006.

[27] B. Granström and D. House, "Audiovisual representation of prosody in expressive speech communication," *Speech Commun.*, vol. 46, pp. 473–484, 2005.

[28] A. Mehrabian, "Framework for a comprehensive description and measurement of emotional states," *Genet Soc. Gen. Psychol. Monogr.*, vol. 121, no. 3, pp. 339–361, 1995.

[29] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Personality*, vol. 11, pp. 273–294, 1977.

[30] A. Mehrabian, "Measures of individual differences in temperament," *Edu. Psychol. Meas.*, vol. 38, pp. 1105–1117, 1978.

[31] P. Gebhard, "ALMA: A layered model of affect," in *4th Int. Joint Conf. Autonom. Agents Multiagent Syst.*, 2005, pp. 29–36.

[32] Z. Y. Wu, H. M. Meng, H. Ning, and C. Tse, "A corpus-based approach for cooperative response generation in a dialog system," in *Proc. 5th Int. Symp. Chinese Spoken Lang. Process.*, Singapore, 2006, vol. 1, pp. 614–626.

[33] "Discover Hong Kong," [Online]. Available: http://www.discoverhongkong.com

[34] J. Alexandersson, Buschbeck-Wolf, M. K. Fujinami, E. M. Koch, and B. S. Reighinger, "Dialogue Acts in VERBMOBIL-2," Univ. Hamburg, DFKI Saarbrucken, Univ. Erlangen, TU Berlin, Germany, Verbmobil Report 226.

[35] M. Nespor and I. Vogel, *Prosodic Phonology*. Dordrecht, The Netherlands: Foris, 1986.

[36] C. Tseng, S. Pin, and Y. Lee, "Speech prosody: Issues, approaches and implications," in *From Traditional Phonology To Modern Speech Processing*, G. Fant, H. Fujisaki, J. Cao, and Y. Xu, Eds. Beijing, China: Foreign Language Teaching and Research Press, 2004, pp. 417–438.

[37] Y. Chao, *A Grammar of Spoken Chinese*. Berkeley, CA: Univ. of California Press, 1968.

[38] H. Kawahara, J. Estill, and 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," in *Proc. Int. Workshop Models and Analysis of Vocal Emissions for Biomedical Applicat.*, 2001.

Zhiyong Wu received the B.S. and Ph.D. degrees in computer science and technology from Tsinghua University, Beijing, China, in 1999 and 2005, respectively.

He has been Postdoctoral Fellow in the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong (CUHK) from 2005 to 2007. He joined the Graduate School at Shenzhen, Tsinghua University, Shenzhen, China, in 2007, where he is currently an Associate Professor. He is also with the Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems. His research interests are in the areas of multimodal multimedia processing and communication, more specifically, audiovisual bimodal modeling, text-to-audio-visual-speech synthesis, and natural language understanding and generation.

Dr. Wu is a member of the Technical Committee of Intelligent Systems Application under the IEEE Computational Intelligence Society and the International Speech Communication Association.

Helen M. Meng (M'99) received the S.B., S.M., and Ph.D. degrees, all in electrical engineering, from the Massachusetts Institute of Technology (MIT), Cambridge.

She has been Research Scientist with the MIT Spoken Language Systems Group, where she worked on multilingual conversational systems. She joined The Chinese University of Hong Kong (CUHK) in 1998, where she is currently a Professor in the Department of Systems Engineering and Engineering Management and Associate Dean of Research of the Faculty of Engineering. In 1999, she established the Human–Computer Communications Laboratory at CUHK and serves as Director. In 2005, she established the Microsoft-CUHK Joint Laboratory for Human-Centric Computing and Interface Technologies, which was upgraded to MoE Key Laboratory in 2008, and serves as Co-Director. She is also Co-Director of the Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems. Her research interest is in the area of human–computer interaction via multimodal and multilingual spoken language systems, as well as translingual speech retrieval technologies.

Prof. Meng serves as the Editor-in-Chief of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. She is also a member of Sigma Xi and the International Speech Communication Association.

Hongwu Yang (A'06) received the M.S. degree in physics from Northwest Normal University, Lanzhou, China, in 1995 and the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 2007.

He is currently an Associate Professor with the College of Physics and Electronic Engineering, Northwest Normal University. His research interests include expressive speech synthesis and recognition, audio content-based information retrieval, and multimedia processing.

Dr. Yang is a member of the Institute of Electronics, Information, and Communication Engineers and a member of the IEEE Signal Processing Society.

Lianhong Cai received the B.S. degree in computer science and technology from Tsinghua University, Beijing, China, in 1970.

She is currently a Professor with the Department of Computer Science and Technology, Tsinghua University. She was Director of the Institute of Human–Computer Interaction and Media Integration from 1999 to 2004. Her major research interests include human–computer speech interaction, speech synthesis, speech corpus development, and multimedia technology. She has undertaken 863 National High Technology Research and Development Program and National Natural Science Foundation of China projects.

Prof. Cai is a member of the Multimedia Committee of Chinese Graphics and Image Society and Chinese Acoustic Society.