An Initial System for Integrated Synthesis of Mandarin, Min-nan, and Hakka Speech

Hung-Yan Gu[1], Yan-Zuo Zhou[1], and Huang-Liang Liau[1]

[1] Department of Computer Science and Information Engineering,
National Taiwan University of Science and Technology, Taipei, Taiwan
{guhy, M9315058, M9215001}@mail.ntust.edu.tw

Abstract. In this study, an integrated speech synthesis system is initially built to synthesize Mandarin, Min-nan, and Hakka speeches. By integration, only a model trained with Min-nan sentences is used to generate pitch-contours for the three languages, same rules are used to generate syllable duration and amplitude values, and a same program module implementing the method, TIPW, is used to synthesize the three languages' speech waveforms. Also, each syllable of a language has just one recorded signal waveform, i.e. no chance of unit selection. Under such a restricted situation, the synthetic speech signals still have a noticeable level of naturalness and signal clarity.

Keywords: cross-lingual speech synthesis, pitch contour, TIPW.

1 Introduction

In Taiwan, there are many languages spoken, including Mandarin, Min-nan, Hakka, and others of smaller population. Mandarin has been most extensively studied than the other languages because it is the official language. However, successful construction of a synthesis model or system for Mandarin does not imply the same modeling method can be directly applied to another language. Developing speech synthesis systems for other languages is strongly aspired because Mandarin is not the mother tongue of most people in Taiwan. Also, some weak languages face the crisis of disappearance.

If systems developed or speech data collected previously can only be applied to Mandarin, then further resource (efforts and money) are still needed to study other languages. Such a situation will become more severe if a corpus-based approach [1], [2] is adopted. In addition, there will be inconsistency in prosody and timbre among independently developed speech synthesis systems for different languages. Therefore, a better approach is to construct a more generalized system that can synthesize not only Mandarin but also Min-nan and Hakka speeches. Such an approach, if successfully realized, can not only save resource but also obtain much higher consistency among the synthesized speeches for different languages. Also, another advantage is that an improvement when made to a system component can immediately benefit all the languages supported.

Mandarin, Min-nan, and Hakka are all syllable prominent languages, and are all tone languages. The number of different syllables (not distinguishing lexical tones) is 405 for Mandarin, 833 for Min-nan, and 783 for Hakka [3]. The numbers of different lexical tones are 5, 7, and 7, respectively, for Mandarin, Min-nan, and Hakka. Since the numbers, 405, 833, and 783, are not large, syllable is commonly chosen as the speech unit for synthesis processing. Actually, in our system, each syllable (tone not distinguished) of a language has only one recorded utterance. That is, no extra units are available to do unit selection, and each syllable’s waveform must be manipulated to synthesize speech signals owning different required prosodic characteristics.

In general, a speech synthesis system can be divided into three subsystems, i.e., (a) text analysis, (b) prosodic parameter generation, and (c) speech waveform synthesis [4], [5]. Our integrated system is also divided into such subsystems. The main processing flow of our system is shown in Fig. 1. The first two processing blocks are for text analysis, the middle two blocks are for prosodic parameter generation, and the last two blocks are for signal waveform synthesis. Our system is an integrated system but not a bundle of three independent systems for the three languages. This is because the program modules in the three subsystems are all shared in synthesizing the three languages’ speeches. For example, the model for pitch-contour parameter generation is shared (or adapted) to Mandarin and Hakka, although it is originally trained with Min-nan training sentences. The explanations for why the program modules can be shared are given in the following sections.

In the subsystem of text analysis, the first block in Fig. 1, ”Text Analysis A”, parse the inputted text to recognize tags and slice the text string into a sequence of Chinese-character or alphanumeric-syllable tokens. Then, each Chinese-character token is tried, in the second block (”Text Analysis B”), to check if it can be looked up in a pronunciation dictionary in order to determine the comprising character’s pronunciation syllable. For the subsystem of prosodic parameter generation, the pitch-contour parameters of a syllable are determined by a mixed model of ANN [6], [7] and SPC-HMM [8] in the third block of Fig. 1. As to the parameters, amplitude and duration, for a syllable, they are determined with a rule-based method [9], [10] in the forth block of Fig. 1. For the subsystem of signal waveform synthesis, a piece-wise linear time-warping function is first constructed in the fifth block of Fig. 1. Then, the method of TIPW (time-proportioned interpolation of pitch waveform, an improved variant of PSOLA) [11] is used in the sixth block of Fig. 1 to synthesize speech signal waveforms. To show that the integrated system has noticeable performance in naturalness and signal clarity, we have set up a web page to provide example synthetic speeches for the three languages (http://guhy.csie.ntust.edu.tw/hmtts/).

2 Text Analysis

In Min-nan and Hakka, there are still many spoken words whose corresponding ideographic words are not known. Therefore, we make a decision that in the inputted text, Chinese characters may be interleaved with syllables spelled in alphanumeric symbols. For example, “cit4 tou5 人” is a Min-nan word whose first two syllables are spelled in alphanumeric symbols. This decision implies that the inputted text must

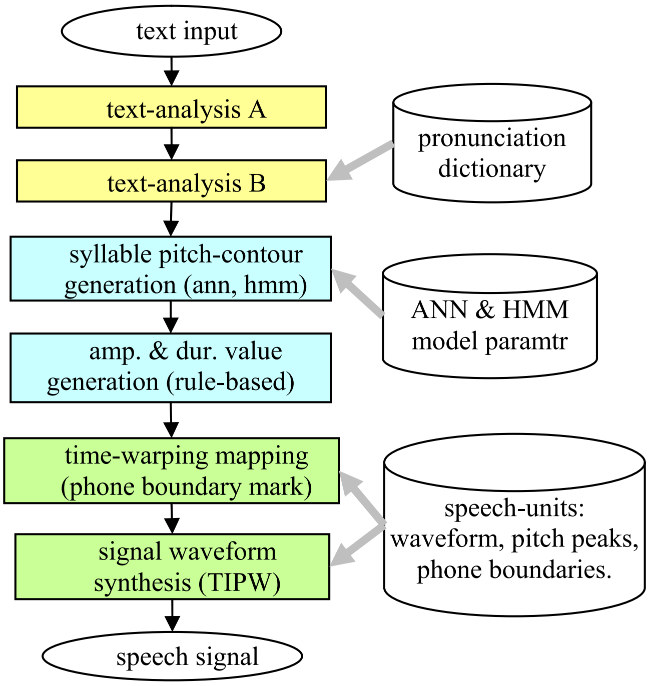
text input
text-analysis A
text-analysis B
pronunciation dictionary
syllable pitch-contour generation (ann, hmm)
ANN & HMM model paramtr
amp. & dur. value generation (rule-based)
time-warping mapping (phone boundary mark)
signal waveform synthesis (TIPW)
speech-units: waveform, pitch peaks, phone boundaries.
speech signal

Fig. 1. Main processing flow.

first be parsed into a sequence of Chinese-character and alphanumeric-syllable tokens. For example, "今天 mai3 ki3" is parsed into "今天", "mai3", and "ki3". The digit at the end of a syllable indicates the lexical tone of the syllable. This parsing processing is executed in the first block of Fig. 1.

In addition, we have defined several kinds of tags to help carry some necessary information. For example, the tag, "@>2", may be placed between two sentences to indicate that the sentences behind the tag will be synthesize to Hakka speech till another language-selection tag is encountered. Such language-selection tag is needed because we intend that the sentences of an article may be alternatively synthesized to different languages' speeches. Another kind of tag is "@>dxxx". This tag may also be placed between two sentences to change the speaking rate of the sentences behind it. The part, "xxx", in the tag represents three decimal digits to specify how many mini-seconds in average a syllable will be synthesized to. In addition to the two tags explained, several other kinds of tags are also defined. The details are listed in Table 1. The parsing of such tags is also executed in the first block of Fig. 1.

After an inputted sentence is parsed into a sequence of tokens, the pronunciation syllables for each Chinese-character token are determined in the second block of Fig. 2. According to the language-selection tag, the corresponding pronunciation dictionaries are looked up to check if the prefix of a token can be found in the dictionaries. The dictionary consisted of longer words is tried before the dictionary

Table 1. Defined tags and their meanings.

Tag symbol	Explanation
@>x	language selection, x may be 0, 1, or 2. 0: Min-nan, 1: Mandarin, 2: Hakka
@>dxxx	speaking rate, syllable average duration in xxx mini-seconds.
@>txxx	average tone height, in xxx Hz.
@>vxxx	vocal track extended (or shrunken) to xxx percents of original length.
<, >	word-constructing tag, e.g., <cit4 tou5>
*	breath-break tag

consisted of shorter words. Currently, we have collected 55,000, 12,000, and 19,000 multi-syllabic words, respectively, for Mandarin, Min-nan, and Hakka. Note that inputted text is usually composed in Mandarin written words. Therefore, dictionary looking-up plays a role of word translation. For example, "今天" (today) in Mandarin is translate to "今仔日", in Min-nan, which is pronounced as, "gin1 a2 rit8". Another example, "筷子"(chopstick) in Mandarin is translated to "箸" which is pronounced as, "di7". These examples also show that the words obtained after translation may have longer or shorter lengths.

After a word is found in a dictionary or a segment of syllables bounded with the tags, "<" and ">", is parsed out, we know the boundaries of a word and its comprising syllables. Then, tone-sandhi rules for the currently selected language can be applied to the comprising syllables of the word. This is executed in the second block of Fig. 1. Note that different languages have very different tone-sandhi rules. For example, in Mandarin, if two adjacent syllables are both of the third tone, then the former one must have its tone changed to the second tone. As another example, consider the tone-sandhi rule in Min-nan that every syllable of a word except the word-final one must have its tone changed to its inflected tone.

3 Prosodic Parameter Generation

The prosodic parameters of a syllable include pitch-contour, duration, amplitude, and leading pause. The generating of prosodic parameter values plays a very important role because it determines the naturalness level of a synthesized speech. Therefore, many efforts had been devoted to investigate models (or methods) for generating of prosodic parameter values [7], [8], [12], [13], [14].

Among these prosodic parameters, pitch-contour is the most important one for higher naturalness level. Therefore, we have spent many efforts in investigating different kinds of models, HMM [8], ANN, and a mixed model of both [15]. In the third block of Fig. 1, a mixed model of HMM and ANN is used to generate pitch-contours. Here, model mixing means taking a weighted sum of two pitch-contours generated by HMM and ANN, respectively. Note that in this study, pitch-contour models, HHM and ANN, are both trained with Min-nan spoken sentences. Then, through tone mapping, we adapt the Min-nan trained and mixed model to generate pitch-contours for Hakka and Mandarin. By such a sharing of pitch-contour model, the efforts for training other languages' pitch-contour models can be saved.

In contrast to pitch-contour, duration and amplitude are thought to be minor factors for naturalness. Hence, only a rule-based method is used in the forth block of Fig. 1 to generate their values.

3.1 Syllable Pitch Contour HMM

A syllable at the beginning of a sentence is usually uttered with higher pitch than the one at the end, i.e., the phenomenon of declining. With respect to this phenomenon, we imagine that there are three prosodic states corresponding to sentence-initial, sentence-middle, and sentence-final. However, we do not know how to assign a sentence's syllables to these states explicitly. Therefore, we imagine these prosodic states are hidden and will simulate them by the hidden states of a left-to-right hidden Markov model. Besides the influence of prosodic states, the lexical tones of a syllable and its adjacent syllables also have strong influences. Therefore, we take into account the lexical-tones of a syllable and its adjacent syllables, and call such an HMM as syllable pitch-contour HMM (SPC-HMM).

The height and shape of a syllable's pitch-contour are mainly influenced by the tone classes of the syllable and its left and right adjacent syllables. Therefore, we decide to combine the *t*'th syllable's lexical tone and pitch-contour VQ (vector quantization) code with its left and right adjacent syllables' lexical tones to define the *t*'th observation symbol, O_t, as

$$\begin{aligned} O_t &= 392 \cdot X_{t-1} + 56 \cdot X_t + 8 \cdot X_{t+1} + V_t \,, \\ &\quad 0 \le X_t \le 6,\ \ 0 \le V_t \le 7 \ . \end{aligned} \tag{1}$$

where X_t is the lexical-tone number of the *t*'th syllable, and V_t is the pitch-contour VQ code of the *t*'th syllable in a training sentence.

Before VQ encoding, the pitch-contour of each syllable from the training sentences is first time normalized and then pitch-height normalized [8]. Time normalization means placing 16 measuring points equally spaced in time. Then, a pitch-contour is represented as a vector of 16 frequency values (in log Hz scale), called a frequency vector. After time normalization, these frequency vectors must be normalized in pitch-height to eliminate the influence of the speaker's emotion in recording time. Totally, we have recorded 643 Min-nan training sentences that are comprised with 3,696 syllables.

Next, consider the generating of pitch-contours by using SPC-HMM. When a sentence is inputted, it will be analyzed first by the text-analysis components. Hence, its pronunciation-syllable sequence is available. Then, we can encode every three adjacent syllables' lexical tones partially (without information of VQ code, V_t) according to Equation (1). Since each tone class has 8 codewords in its pitch-contour VQ codebook, each syllable of the sentence has 8 possible encoded observation symbols corresponding to it. Therefore, in the synthesis phase (or testing phase), besides the time and state axes, a third axis, i.e. index to the 8 possible observation symbol candidates, must be added. Then, the conventional two-dimensional (time and state) DP (dynamic programming) algorithm for speech recognition is extended to three-dimensional DP algorithm and used to search the most probable path [8]. According to the path found, pitch-contour VQ codes of the syllables can then be decoded.

3.2 Syllable Pitch Contour ANN

The architecture of the artificial neural network used here is as shown in Fig. 2. It is designed to be a recurrent type ANN in order to have prosodic state kept internally. The input layer of the ANN has 8 ports (28 bits) to receive contextual parameters. For the hidden and recurrent hidden layers, the numbers of nodes are both set to be 30 according to experiment results. After a syllable's contextual parameters are inputted and processed, a pitch contour represented as a 16 dimensional frequency vector is outputted at the output layer.

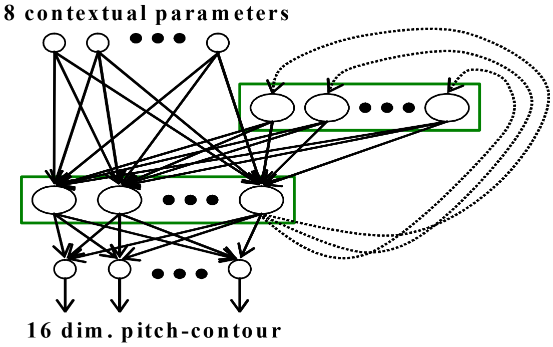

Fig. 2. The architecture of the ANN studied here.

Here, the contextual parameters, i.e. the inputs to the ANN, are appropriately selected to provide essential contextual information and to lower the quantity of required training sentences. In details, the contextual parameters are as listed Table 2. Because there are totally seven lexical tones in Min-Nan, 3 bits are enough to represent them. The numbers of different syllable initials and finals are respectively 18 and 61. Hence, 5 and 6 bits are used respectively to represent current syllable's initial and final types. As to the parameter of the previous syllable's final, we first group the 61 possible finals into 12 classes, and use only 4 bits to represent the 12 final classes. Similarly, we first group the 18 possible initials into 6 classes, and use only 3 bits to represent the 6 initial classes for the next syllable's initial. Grouping is made here because the quantity of recorded training sentences is not large enough to let the ANN learn the influences of the detailed combinations of current syllable and previous (or next) syllable. Syllable initial and final classes grouped here are detailed in Table 3 and 4, respectively. The last item in Table 2, i.e. time-progress index, is intended to carry timing information. If the current syllable is the *k*'th syllable of a sentence of *N* syllables in length, then time-progress index is defined as *k*/*N*.

Table 2. Contextual parameters.

Items	Tone of previous syllable	Final class of previous syllable	Tone of current syllable	Initial of current syllable	Final of current syllable	Tone of next syllable	Initial class of next syllable	Time progress index
Bits	3	4	3	5	6	3	3	1

Table 3. Syllable-initial classes. (General Phonetic Symbol System)

Classes	1	2	3	4	5	6
Initials	(null), m, n, l, r, ng, q, v	h, s	b, d, g	z	c	p, t, k

Table 4. Syllable-final classes. (General Phonetic Symbol System)

Classes	Finals	Classes	Finals
1	(null)	7	-i, -u, -ui, -iu
2	-a, -ia, -ua	8	-ing, -eng, -in, -un, -en
3	-o, -io, -ior	9	-ang, -iang, -uang, -ong, -iong, -an, -uan
4	-er, -ier	10	-am, -iam, -im, -om
5	-e, -ue	11	-ah, -eh, -ih, -oh, -uh, -auh, -erh, -iah, -ierh, -ioh, -uah, -ueh
6	-ai, -uai, -au, -iau	12	-ap, -iap, -ip, -op, -at, -et, -it, -uat, -ut, -ak, -iak, -ik, -iok, -ok

3.3 Adaptation of Pitch Contour Model

Here, the pitch-contour model trained with Min-nan sentences is used as the working model. We develop an adaptation method that will not change any parameter' values of the working model. In details, the lexical tone sequence, $X_1X_2, \ldots, X_n$, collected from a target (Mandarin or Hakka) language's sentence are first mapped to the working language's lexical tone sequence, $Y_1Y_2, \ldots, Y_n$. The mapped lexical tones are then used as the input to the working model. And the generated pitch-contours, $R_1R_2, \ldots, R_n$, by the working model are treated as the output of the adapted model for the target language.

The reasons why the adaptation method can work are explained in details in our previous paper [16]. In brief, slight differences in frequency-height or boundary-part shape between two pitch-contours need not be worried for correct recognition of the carried lexical tone. And we think it is reasonable to approximate a pitch-contour shape in a target language with a shape class trained in the working language that is of similar shape trend in the central part. Note that each lexical tone of the working language is usually trained to have several representative shape classes, e.g. 8 code-words in each tone's VQ codebook. Hence, for a pitch-contour shape in a target language, we can select from the shape classes trained for the mapped lexical tone to pick out the one that is most similar. Then, the possible decrease in naturalness due to differences in frequency-height and boundary-part shape can be minimized.

The mapping from Hakka tones to Min-Nan tones is listed in Table 5. This mapping can be said to be a nice one-to-one mapping because both have the same number of lexical tones, and for each lexical tone of Hakka we can find a lexical tone in Min-Nan that has almost same pitch-contour shape.

Table 5. Tone mapping from Hakka to Min-nan.

Hakka tone number	1	2	3	4	5	7	8
Mapped Min-Nan tone number	2	5	3	8	1	7	4

The mapping from Mandarin tones to Min-Nan tones is listed in Table 6. Three of the Mandarin lexical tones, i.e. high-level, rising, and falling, are also existent as Min-Nan lexical tones. Besides these three tones, the low-level tone of Min-nan and the low-dipping tone of Mandarin are perceived as the same tone. Therefore, the low-dipping tone of Mandarin can be mapped to the low-level tone. As to neutral tone of Mandarin, it has shorter duration than the other tones. This contrast also exists in Min-nan, i.e. both abrupt tones have shorter durations. In addition, the low-abrupt tone of Min-nan has low pitch-height as the neutral tone has. Hence, the neutral tone of Mandarin can be mapped to low-abrupt tone of Min-nan.

Table 6. Tone mapping from Mandarin to Min-nan.

Mandarin tone number	1	2	3	4	5
Mapped Min-Nan tone number	1	5	3	2	4

4 Signal Waveform Synthesis

In our system, each syllable of a language has only one utterance recorded in level tone. Therefore, an original syllable waveform must be manipulated to synthesize target syllable waveforms of different prosodic characteristics. The synthesis method used here is TIPW [11]. TIPW is an improved variant of PSOLA, i.e. the effects of chorus and reverberation are largely reduced. Besides, TIPW is capable of adjusting vocal track length through re-sampling [17].

Originally, TIPW is developed for synthesizing Mandarin speech. Thus, it does not support the synthesis of abruptly changing signal waveform that is found at the ending portion of an abrupt-tone syllable, e.g. /zit8/. Nevertheless, abrupt-tone syllables are very frequently used in Min-nan and Hakka. One method to overcome this difficulty is to treat the ending portion of an abrupt-tone syllable as a stop consonant. Then, the same method as used in synthesizing stop consonant at syllable initial portion can also be adopted to solve this problem.

In addition, we have made another improvement to TIPW. This improvement significantly increases the fluency of the synthesized speech. In ordinary speech synthesis system, the subsystem of prosodic-parameter generating only determines the duration value, *Es*, of a syllable to be synthesized. The detailed dividing of syllable duration, *Es*, to its comprising phonemes is however not controlled by the prosody subsystem. Furthermore, the subsystem of signal waveform synthesis usually extends (or shrinks) the original speech waveform to an intended time length in a linear way. According to our study, linear extending (or shrinking) of time length is an important factor that decrease much the fluency of a synthesized speech.

Let us consider the example syllable, /man/. Suppose that in its original recorded waveform, the three phonemes, /m/, /a/, and /n/, occupy *Dm*, *Da*, and *Dn* mini seconds, respectively, and *Ds* = *Dm* + *Da* + *Dn*. A phenomenon that can be observed is that the ratio, (*Dm*+*Dn*)/*Ds*, will become smaller when /man/ is uttered within a sentence instead of being uttered isolatedly. Currently, we study to simulate this phenomenon with a simple method. In the further, we will study it with a more systematic method. The method used here is as depicted in Fig. 3. That is, a piecewise linear function is used to map the time-axis of a synthetic syllable waveform to the time-axis of its original waveform. In Fig. 3, the symbols, *Em*, *Ea*, and *En* represent the time lengths of the three phonemes, /m/, /a/, and /n/ in the synthesized waveform while *Dm*, *Da*, and *Dn* represent these three phonemes' time lengths in the original syllable waveform. In our system, the values of *Em*, *Ea*, and *En* is determined with the following procedure:

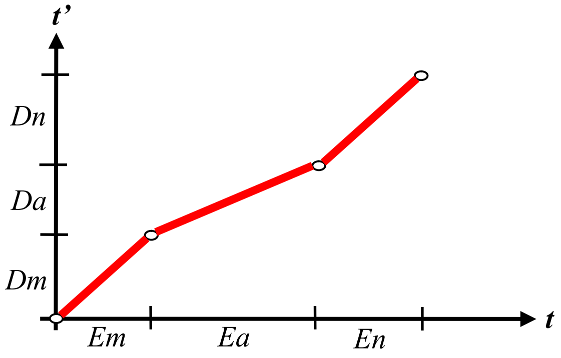

Fig. 3. Piece-wise linear mapping function.

```
r = 0.7;
do {
   Em = (Dm/Ds) * r * Es;
   En = (Dn/Ds) * r * Es;
   Ea = Es - Em - En;
   if(Ea > Es*0.4) break;
   r = r - 0.05;
} while(r>=0.2);
```

If the structure of a syllable is the same as /san/ or /an/, i.e. without voiced initial consonant, then the values of *Dm* and *Em* can be set directly to zero. Similarly, if the structure of a syllable is the same as /ma/, i.e. without voiced ending consonant, then the values of *Dn* and *En* can be set directly to zero. Apparently, to apply the procedure given above, the boundary point between two adjacent phonemes must be labeled beforehand in order to compute the values of *Dm*, *Da*, and *Dn*, and to construct the mapping function.

5 Experiments and Results

After the integrated system is implemented and ready to run, we then start to do perception tests. The first issue we are concerned is the naturalness levels of Mandarin and Hakka pitch-contours generated by the Min-nan trained pitch contour model. Therefore, short articles written in Mandarin and Hakka are fed as inputs to our system. The outputted speech signals are then played to each of the ten persons who participates the tests. Initial results show that the lexical tones in the synthetic Mandarin and Hakka speeches can be correctly recognized. As to naturalness level, the synthetic Hakka speech is perceived to be slightly more natural than the synthetic Mandarin speech. This is because the pitch-contours generated for Mandarin is perceived to have slightly strange accent.

Another issue we are concerned is the fluency of the synthetic speech. Therefore, two synthesis conditions are set to synthesize a short article into two Mandarin speeches. In the first condition, the mapping function (used in waveform synthesis) between synthetic and original time axes is forced to be linear while in the second condition, the mapping function adopted is as showed in Fig. 3. The two synthetic speeches are then played to each of the participating persons to evaluate their fluency. Initial results show that the fluency of the synthetic speech under the second condition is significantly better than the one under the first condition.

Acknowledgments. This study is supported by National Science Council under the contract number, NSC 94-2218-E-011-007.

References

1. Chou, F.-C.: Corpus-based Technologies for Chinese text-to-Speech Synthesis. Ph.D. Dissertation, Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan (1999)
2. Chu, M., *et al.*: Microsoft Mulan - a bilingual TTS system. ICASSP'03, Vol. 1 (2003) I264-I267
3. Yu, B.-C., Syu, Z.-C., Wu, C.-N.: General Phonetic Symbol System for Languages in Taiwan. Nan-Tien Book Company, Taipei (1999)
4. Shih, C., Sproat, R.: Issues in Text-to-Speech Conversion for Mandarin. Computational Linguistics & Chinese Language Processing, Vol. 1, No. 1 (1996) 37-86
5. Wang, R.-H.: Overview of Chinese Text-to-Speech System. ISCSLP'98, Singapore (1998)
6. Lee, S. J., Kim, K. C., Jung, H. Y., Cho, W.: Application of Fully Recurrent Neural Networks for Speech Recognition. ICASSP'91, (1991) 77-80
7. Chen, S. H., Hwang, S. H., Wang, Y. R.: An RNN-based Prosodic Information Synthesizer for Mandarin Text-to-Speech. IEEE trans. Speech and Audio Processing, Vol. 6, No.3 (1998) 226-239
8. Gu, H. Y., Yang, C. C.: A Sentence-Pitch-Contour Generation Method Using VQ/HMM for Mandarin Text-to-speech. ISCSLP'2000, Beijing (2000) 125-128
9. Chiou, H. B., Wang, H. C., Chang, Y. C.: Synthesis of Mandarin speech based on hybrid concatenation. Computer Processing of Chinese and Oriental Languages, Vol. 5 (1991) 217-231
10. Shiu, W.-L.: A Mandarin Speech Synthesizer Using Time Proportioned Interpolation of Pitch Waveform. Master Thesis, National Taiwan University of Science and Technology, Taipei (in Chinese) (1996)
11. Gu, H.-Y., Shiu, W.-L.: A Mandarin-Syllable Signal Synthesis Method with Increased Flexibility in Duration, Tone and Timbre Control. Proc. Natl. Sci. Counc. ROC(A), Vol. 22, No.3 (1998) 385-395
12. Lee, L.-S., Tseng, C.-Y., Hsieh, C.-J.: Improved Tone Concatenation Rules in a Formant-based Chinese Text-to-Speech System. IEEE trans. Speech and Audio Processing, Vol. 1 (1993) 287-294
13. Wu, C.-H., Chen, J.-H.: Automatic Generation of Synthesis Units and Prosodic Information for Chinese Concatenative Synthesis. Speech Communication, Vol. 35. (2001) 219-237
14. Yu, M. S., Pan, N. H., Wu, M. J.: A Statistical Model with Hierarchical Structure for Predicting Prosody in a Mandarin Text-to-Speech System. ISCSLP'02*,* Taipei (2002) 21-24
15. Gu, H.-Y., Huang, W.: Min-Nan Sentence Pitch-contour Generation: Mixing and Comparison of Two Kinds of Models. ROCLING'05, Tai-Nan, Taiwan (in Chinese) (2005) 213-225
16. Gu, H.-Y., Tsai, H.-C.: A Pitch-Contour Model Adaptation Method for Integrated Synthesis of Mandarin, Min-Nan, and Hakka Speech. 9'th IEEE International Workshop on Cellular Neutral Networks and their Applications, Hsinchu, Taiwan, (2005) 190-193
17. Gu, H.-Y.: Signal Resampling in Speech Synthesis. 5'th World Multi-conference on Systemics, Cybernetics and Informatics (Orlando, USA), Vol. VI, (2001) 521-525