A New Model-based Mandarin-speech Coding System

Chen-Yu Chiang[1], *Jyh-Her Yang*[1], *Ming-Chieh Liu*[1], *Yih-Ru Wang*[1], *Yuan-Fu Liao*[2] *and Sin-Horn Chen*[1]

[1]Institute of Communication Engineering, National Chiao Tung University, Taiwan
[2]Department of Electronic Engineering, National Taipei University of Technology, Taiwan

gene.cm91g@nctu.edu.tw, {neil.yang0204, ufo0241}@gmail.com, yrwang@mail.nctu.edu.tw, yfliao@ntut.edu.tw, schen@mail.nctu.edu.tw

Abstract

In this paper, a new model-based Mandarin-speech coding system is proposed. It employs a prosody-enriched ASR with a hierarchical prosodic model (HPM) to generate from the input speech enriched transcriptions, including linguistic features, prosodic tags and spectral parameters in the encoder. By sending these features to the decoder, we can first reconstruct the prosodic-acoustic features of syllable pitch contour, syllable duration, syllable energy level, and inter-syllable pause duration by HPM using the linguistic features and prosodic tags; and then combined with spectral parameters to reconstruct the input speech signal by an HMM-based speech synthesizer. Experimental results show that the reconstructed speech has good quality at a low data rate of 543 bits/s.

Index Terms: model-based speech coding, prosody-enriched ASR, enriched transcriptions, hierarchical prosodic model.

1. Introduction

Speech coding is conventionally performed on a sample-based approach to take advantage of the inter-sample redundancy [1], or on a frame-based approach to separately manipulate on spectral and excitation features [2]. By virtue of advances in automatic speech recognizer (ASR) and HMM-based speech synthesis technologies [3], some phonetic/segment-based vocoders were also reported [4,5]. Usually, only local articulation information is used in the above-mentioned studies. If we can use higher level linguistic feature, such as phone/syllable/word, and/or prosodic features, such as syllable pitch contour, duration, energy level, then the coding efficiency can be further improved. Besides, post processing on those linguistic or prosodic features can conduct to some interesting applications. For examples, changing the speaking rate can be accomplished by adjusting the prosodic features; while voice conversion can be realized by adjusting both the spectral and prosodic features of frame/syllable/word. One way to extract high-level linguistic features and prosodic features from input speech signal is by using a high-performance ASR. Due to the conceptual similarity of the approach to the model-based image coding, we call it the model-based speech coding approach.

In this paper, a new model-based Mandarin speech coding system is proposed. A high-performance prosody-enriched ASR is incorporated in the coding system to provide enriched transcription and segmentation information of the input speech. Information decoded includes linguistic feature strings of base-syllable, tone, word, part-of-speech (POS) and punctuation mark (PM), as well as prosodic tag sequences of syllable prosodic state and inter-syllable break type. Assisted with these linguistic and prosodic features, the coding efficiency can be greatly improved.

The remainder of the paper is organized as follows. Section 2 presents the proposed Mandarin-speech coding system in detail. Section 3 discusses the performance evaluation of the system. Some discussions and conclusions are given in the last section.

2. The Proposed Coding System

Fig. 1 shows a schematic diagram of the proposed Mandarin-speech coding system. In the encoder, input speech signal is firstly decoded by a prosody-enriched Mandarin ASR system (PE-ASR) [6,7] with an HMM-based acoustic model (AM), a factored language model (FLM) [8] and a hierarchical prosodic model (HPM) [9]. Three types of information are transcribed by the decoding. One is linguistic features including strings of base-syllable, tone, word, POS and PM. Another is prosodic features including tag sequences of syllable prosodic state and inter-syllable break type. It is worth to note that these two prosodic tag sequences can be used to form a hierarchical prosody structure of the input speech. The other is the segmentation information of various levels from HMM state to word.

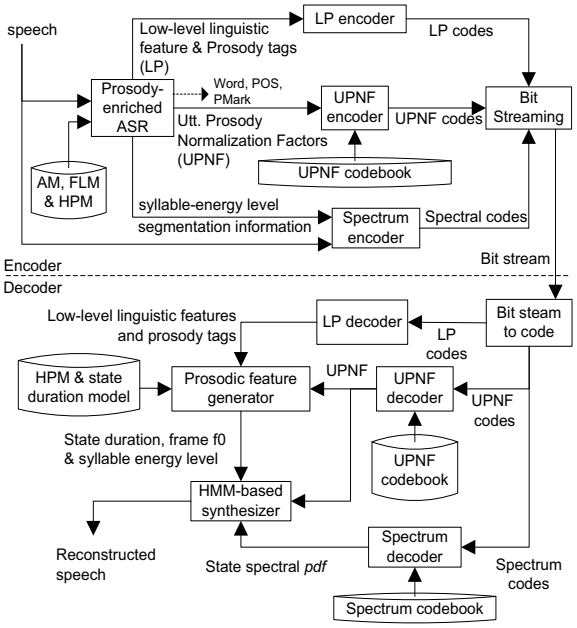

Fig. 1: *A schematic diagram of the proposed speech system.*

By using some low-level linguistic features and prosodic tags (LP), we can reconstruct prosodic-acoustic features, including syllable pitch contour, syllable duration, syllable energy level, and inter-syllable pause duration with the help of HPM. So, we only need to encode those LP features for prosody reconstruction in the decoder. It is noted that prosodic features used in PE-ASR are pre-normalized by speaker-level (training phase) or utterance-level (test phase) mean and variance. Therefore, an additional utterance prosody

normalization factor (UPNF) encoder is required for encoding these prosody normalization factors. By using the HMM-state segmentation information, we can extract state-based spectral features and encode them by vector quantization (VQ).

In the decoder, we first use the decoded LP features to reconstruct the four prosodic-acoustic features by HPM whose parameters are sent to the decoder in advance as side information. We then use base-syllable type and syllable duration to predict state durations by a state duration model. Lastly, by using the decoded state spectral features, the reconstructed prosodic-acoustic features, and the predicted state durations, an HMM-based speech synthesizer generates the output speech.

In the following subsections, we discuss the encoder and the decoder in more detail.

2.1. The Speech Encoder

As shown in Fig. 1, the speech encoder is composed of four parts including a PE-ASR [6,7], an LP encoder, a UPNF encoder, and a spectrum encoder. The PE-ASR system is a sophisticated speech recognizer developed previously [6,7]. Fig. 2 displays its functional block diagram. It is a two-stage system to firstly use an AM and a bigram LM to generate a word lattice in the first stage decoding, and to then use an FLM [8] and an HPM [9] to finely decode from the word lattice the best linguistic sequences (i.e. base-syllable, tone, word, POS and PM) and their corresponding segmentation information, as well as prosody tag sequences (i.e. prosodic states and break types) that represent a hierarchical prosody structure of the input utterance. The AM is a syllable-based HMM model. It models each of 411 base-syllables as an 8-state left-to-right HMM. The FLM is an extension of the conventional trigram model to additionally consider POS and PM aside from word. The HPM consists of various prosodic sub-models to describe the relationship of prosodic tags, prosodic-acoustic features, and linguistic features.

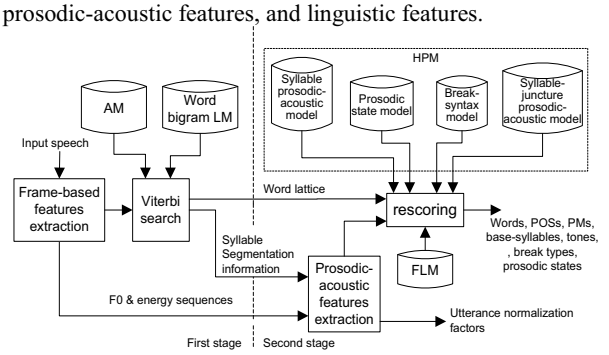

Fig. 2: *The prosody-enriched ASR system* [7]

Four sub-models of the HPM are involved in the coding process. They include three syllable prosodic-acoustic models, which are used to describe the variations of syllable pitch contour, duration and energy level, and one prosodic-acoustic model which describes the variation of syllable-juncture pause duration influenced by some linguistic features. For syllable pitch contour, it is formulated as an additive model:

$$sp_n = sp_n^r + \beta_{t_n} + \beta_{p_n} + \beta^f_{B_{n-1},t^n_{n-1}} + \beta^b_{B_n,t^{n+1}_n} + \mu_{sp} \quad (1)$$

where sp_n is a vector of four orthogonally-transformed parameters representing the observed log-F0 contour of syllable n [10]; sp_n^r is the residual of modeling sp_n; β_{t_n} and β_{p_n} are the affecting patterns (APs) for tone t_n and prosodic state tag p_n, respectively; $\beta^f_{B_{n-1},t^n_{n-1}}$ and $\beta^b_{B_n,t^{n+1}_n}$ are the forward and backward coarticulation APs contributed from syllable $n-1$ and syllable $n+1$, respectively; and μ_{sp} is the global mean of pitch vector. Here, B_n is the break tag after syllable n. Similarly, syllable duration and energy level are modeled as

$$sd_n = sd_n^r + \gamma_{t_n} + \gamma_{s_n} + \gamma_{q_n} + \mu_{sd} \quad (2)$$

$$se_n = se_n^r + \omega_{t_n} + \omega_{f_n} + \omega_{r_n} + \mu_{se} \quad (3)$$

where $\gamma_{t_n}/\omega_{t_n}$, γ_{s_n}, ω_{f_n} and $\gamma_{q_n}/\omega_{r_n}$ are the APs of tone t_n, base-syllable s_n, final type f_n, and prosodic state tags q_n/r_n; and μ_{sd} and μ_{se} are global means. To reconstruct these three prosodic-acoustic features using the three sub-models in the decoder, we need to encode and transmit low-level linguistic features of tone, base-syllable and final types as well as prosodic features of break type and prosodic state tags. Besides, all affecting patterns are sent as side information. It is noted that we neglect the coding of the residuals because they all have small variances.

The fourth sub-model describes the variation of inter-syllable pause duration by break-dependent decision trees (BDTs). For each break type, a decision tree is used to determine the *pdf* of pause duration according to linguistic features. For reconstructing the pause duration, we needs to send the information of the break tag and the residing leaf node of the associated decision tree for each inter-syllable juncture to the decoder. All *pdf*s of leaf nodes in these seven decision trees are also sent to the decoder as side information. Table 1 shows the bit assignment of the encodings of these low-level linguistic features of tone, base-syllable and final types, prosodic tags of prosodic state and break type, and leaf nodes of BDTs. Notice that the BDT is constructed for each break type, and each BDT has different number of leaf nodes. Therefore, the bit length is variable for each given known break type.

Table 1: *Bit assignment for encoding linguistic features and prosody tags.*

Symbol	# of symbol	bit
Lexical tone t_n	5	3
Base-syllable type s_n	411	9
Pitch prosodic state p_n	16	4
Duration prosodic state q_n	16	4
Energy prosodic state r_n	16	4
Break type B_n	7	3
BDT leaf node index T_n for B0, B1, B2-1, B2-2, B2-3, B3, B4	5/7/3/2/4/3/1	3/3/2/1/2/2/0
Total bits per syllable (maximum)		30

For avoiding taking care of the speaker/utterance variability of prosodic-acoustic features in HPM, they are pre-normalized. For syllable pitch contour, a scheme of frame-based F0 value normalized by speaker-level (training phase) or utterance-level (test phase) mean and variance is adopted; while for both syllable duration and syllable energy level, they are simply normalized by their corresponding speaker-/utterance-level means and variances. These normalization factors are needed to be encoded and sent to the decoder. In this study, they are scalar-quantized independently by the UPNF encoder. Their codebooks are also sent to the decoder as side information.

Since we want to use the HMM-based speech synthesizer in the decoder to generate the output speech, we extract 25-dimensional mel-generalized cepstral (MGC) [11] vector including the zero-th coefficient for each 25ms frame with 5ms shift. Blackman window is used in the feature extraction. Besides, delta and delta-delta MGCs are also extracted. In the training phase, we calculate the *pdf* parameters (i.e., mean and

variance) of each MGC coefficient for each HMM state using the training data with the time-aligned segmentation information provided by the PE-ASR system. In the test phase, we first calculate the mean vector of 25-dimentional MGC vectors for each state segment and then subtract the mean MGC vector of the corresponding state of the recognized base-syllable to obtain a residual vector. Lastly, we encode all state-based residual vectors by vector quantization (15 bit for each state). Both the *pdf* parameters of all HMM states and the VQ codebooks are sent to the decoder as side information. It is noted that the energy coefficient in each state MGC vector is pre-normalized by the energy level of the associated syllable.

Table 2 summarizes the side information of the coding system.

Table 2: *Side information of the proposed coding system*

Type	parameter #
Lexical tone APs: $\beta_t / \gamma_t / \omega_t$	5/5/5
Coarticulation APs: $\beta^f_{B,t} / \beta^b_{B,t}$	180/180
Prosodic state APs: $\beta_p / \gamma_q / \omega_r$	16/16/16
Global mean APs: $\mu_{sp} / \mu_{sd} / \mu_{se}$	1/1/1
Base-syllable type and final type APs: γ_s / ω_{f_n}	411/40
BDT leaf node mean: $\mu^{pd}_{T_n}$	25
Spectrum codebook	1056
MGC *pdf*s of all HMM states	26304
Normalization factor codebooks	384
Total	28646

2.2. The Speech Decoder

The task of the speech decoder is to reconstruct speech signal by using the decoded linguistic, prosodic and spectral parameters. As shown in Fig. 1, the speech decoder consists of five parts including the LP decoder, the UPNF decoder, the spectrum decoder, the prosodic-acoustic feature generator, and an HMM-based speech synthesizer [12]. The LP decoder generates low-level linguistic features and prosody tags by looking up tables. The spectrum decoder uses the spectrum codebook to generate the output spectral features of each state from the input codeword index. The prosodic feature generator reconstructs the three prosodic-acoustic features and pause duration by HPM using the decoded low-level linguistic features and prosody tags. These three prosodic-acoustic features are de-normalized by using the decoded utterance-level factors. After obtaining syllable duration, we then predict state durations. Lastly, the HMM-based speech synthesizer reconstructs the input speech signal by using the state spectral features, state duration and the associated prosodic-acoustic features.

In state duration prediction, we assume that the state duration is normally distributed and affected by base-syllable type s_n, i.e

$$P(d_{n,c}|s_n,c)=N(d_{n,c};\mu_c^{s_n},\sigma_c^{s_n}) \quad (4)$$

where $d_{n,c}$ denotes the duration of the c-th state in the n-th syllable. Given the reconstructed syllable duration sd_n, state durations of the syllable can be obtained by maximizing the summed log likelihood, i.e.

$$d^*_{n,1}...d^*_{n,C} = \arg\max_{d_{n,1}...d_{n,C}} \sum_{c=1}^{C} \log P(d_{n,c}\,|\,s_n,c) \quad (5)$$

under the constraint

$$sd_n = \sum_{c=1}^{C} d_{n,c} \quad (6)$$

The resulting state duration can be obtained by

$$d_{n,c} = \mu_c^{s_n} + \rho \cdot (\sigma_c^{s_n})^2 \quad (7)$$

where

$$\rho = \left(sd_n - \sum_{c=1}^{C} \mu_c^{s_n} \right) / \left(\sum_{c=1}^{C} (\sigma_c^{s_n})^2 \right) \quad (8)$$

3. Performance Evaluation

The proposed model-based Mandarin-speech coding system was evaluated on a large Mandarin read speech database TCC300 [13]. The database consists of two sets: 103-speaker short sentential utterances (Set A) and 200-speaker long paragraphic utterances (Set B). The database was collected for Mandarin ASR. Set A was designed to consider the phonetic balance of Mandarin speech, while Set B was designed to additionally consider the usage for prosody study. The database was divided into a training set (about 90%, 274 speakers, 23 hours) and a test set (about 10%, 29 speakers, 2.43 hours). A set of 411 8-state base-syllable HMM models was generated from the training set by HTK 3.4 [14] with the MMIE criterion [15]. The acoustic feature vector is composed of 12 MFCCs and their delta and delta-delta terms, one delta energy and one delta-delta energy. For testing the PE-ASR system, the Set B part of the test set was used. The test subset contained 226 utterances of 19 speakers with length about 2 hours. The total number of words in the test subset is 14993. All testing data were long utterances with average length of 117.2 syllables.

A text corpus was employed to train both the word-bigram LM and the FLM which were used, respectively, in the first- and second-stage speech decodings. The corpus contained in total about 139 million words. A 60,000-word lexicon was also constructed based on word frequency.

Table 3 shows the performance of the PE-ASR system. Word, character, and base-syllable error rates of 20.5%, 14.3%, and 9.9% were achieved, respectively. This performance is very good as compared with most conventional HMM-based ASR methods. Since syllable insertion and deletion errors were expected to cause more serious degradation on the coding performance, we also list them in Table 3. As seen from the table, both of them are small.

Table 3: *The performance of the PE-ASR (%).*

WER	CER	SER	Syll-INS	Syll-DEL	Syll-SUB
20.5	14.3	9.9	0.55	0.83	8.5

We then examined the performance of the coding system. Two cases were examined. One was the inside test in which both the speech utterance and the associated text were given. In this case, we first segmented the speech by time-alignment using the AM, and then labeled the prosodic tags automatically by the HPM. We then performed the encoding and decoding operations to reconstruct the speech. The other case was the outside test in which only the speech utterance was given. This is the case of the proposed coding system discussed in Section 2.

Table 4 shows the root-mean-square errors (RMSE) of the reconstructed four prosodic features. Here, all six utterance-level normalization factors were encoded using 6-bit scalar quantizers. Table 5 shows the RMSE of the reconstructed pause duration for different break types. Since major breaks like B3 and B4 are tolerant of larger errors, the performance was good. The average bit rates were 528 and 543 bits/s for the inside and outside tests, respectively. These data rates are

low. Fig.3 shows an example of the reconstructed prosodic features of an utterance of the outside test. As shown in the figure, most reconstructed prosodic features were close to their reference values.

Table 4: *The RMSE of the reconstructed prosodic features*

	F0 (Hz)	Syllable duration (ms)	Syllable energy level (dB)	Pause duration (ms)
Inside test	11.4	18.4	0.52	73.8
Outside test	14.7	16.8	0.20	75.6

Table 5: *The RMSE (ms) performance of the reconstructed pause duration with respect to different break types.*

	B0	B1	B2-1	B2-2	B2-3	B3	B4
Inside	19.3	26.5	75.6	149.2	35.0	177.9	312.9
Outside	12.4	17.1	88.3	178.4	39.6	176.9	292.7

Table 6: *Bit rates for inside and outside tests*

		Average	Max	Min
Inside	prosody	104.56	163.79	42.23
	spectral	423.73	661.90	178.07
outside	prosody	107.55	147.20	78.00
	spectral	435.06	594.44	318.05

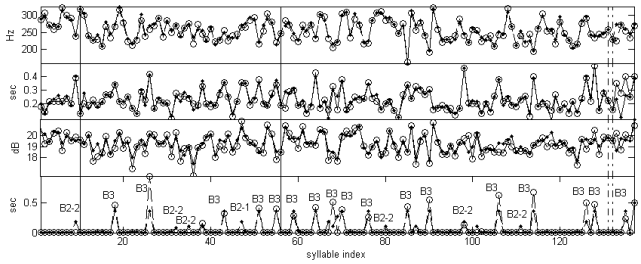

Fig. 3: An example of the reconstructed prosodic features of an utterance. From top to bottom: syllable pitch mean, syllable duration, syllable energy level, and pause duration. (open circle: reference, dot: recognition result, solid line: deletion, dash dot line: insertion)

Lastly, an informal listening test was performed. Generally, all reconstructed speeches sounded good. The effects of recognition errors were not serious. Most substitution, deletion, and insertion errors were slightly perceptible. This mainly resulted from encoding and sending the spectral features to the decoder.

4. Discussions and Conclusions

A model-based Mandarin-speech coding system has been discussed in this paper. It differs from the conventional speech coding system on using a prosody-enriched ASR in the encoder to extract high-level linguistic and prosodic features to assist in improving the coding efficiency. Experimental results showed that high-quality reconstructed speech can be obtained at a low data rate of 543 bits/s.

Another advantage of the proposed coding system can be found. By properly adjusting the prosodic features, we may modify the prosody of the reconstructed speech, e.g. changing the speech rate.

The proposed coding system can also operate on another two modes. One is the case of knowing both the speech signal and the associated text. This case has been examined as the inside test discussed in Section 3. An application of the mode is the speech coding of story readings in an electronic book. Prosody modification will be the most attractive feature of the application. The other mode is the case of low-rate speech coding without transmitting the spectral parameters. A text-to-speech system, such as the HTS [16] can be used to generate spectral parameters of a standard voice for their substitutions by using the recognized text sent from the encoder. In this case, we can keep the prosody of the input speech but losing the speaker identity.

5. Acknowledgements

The work was supported by the NSC, Taiwan, under the project with contracts NSC 99-2221-E-009-009-MY3 and . 98-2221-E-009-075-MY3. The authors would like to thank the ACLCLP for providing the TCC300 Corpus, NTCIR, and Sinca Corpus.

6. References

[1] Jayant, N. S., "Digital coding of speech waveforms: PCM, DPCM, and DM quantizers," Proceedings of the IEEE , vol.62, no.5, pp. 611- 632, May 1974.

[2] Schroeder, M., Atal, B., "Code-excited linear prediction(CELP): High-quality speech at very low bit rates," Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85. , vol.10, no., pp. 937- 940, Apr 1985.

[3] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura T., "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis", Proc. of Eurospeech, pp.2347-2350, Sept. 1999.

[4] Hoshiya, T., Sako, S., Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T., "Improving the performance of HMM-based very low bitrate speech coding", Proc. ICASSP, pp.800–803, 2003.

[5] Tokuda, K., Masuko, T., Hiroi, J., Kobayashi, T. and Kitamura, T., "A very low bit rate speech coder using HMM-based speech recognition/synthesis technique," in Proc. ICASSP, 1998.

[6] Yang, J. H. Liu, M. J., Chang H. H. Chiang, C. Y., Wang, Y. R. and Chen, S. H., "Enriching Mandarin speech recognition by incorporating a hierarchical prosody model", accepted and to appear in ICASSP 2011.

[7] Chen, S. H., Yang, J. H., Chiang, C. Y., Liu, M. C., and Wang, Y. R., "A New Prosody-Assisted Mandarin ASR System", submit to Trans. on IEEE Audio, Speech, & Language Processing.

[8] Bilmes, J. A. and Kirchhoff, K., "Factor language models and generalized parallel backoff," in Proc. of HLT/NACCL, 2003, pp. 4-6.

[9] Chiang, C. Y., Chen, S. H., Yu, H. M. and Wang, Y. R., "Unsupervised joint prosody labeling and modeling for Mandarin speech," Journal of the Acoustic Society of America, 125, No. 2, pp.1164-1183, Feb. 2009.

[10] Chen, S. H. and Wang, Y. R., "Vector quantization of pitch information in Mandarin speech," IEEE Transactions on Communications, vol. 38, no. 9, pp. 1317-1320, September 1990.

[11] Tokuda, K., Masuko, T., Kobayashi, T. and Imai, S., "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation," in Proceedings of the International Conference on Spoken Language Processing (ICSLP '94), pp. 1043–1046, Yokohama, Japan, September 1994.

[12] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., "Speech parameter generation algorithms for HMM-based speech synthesis", Proc. of ICASSP, pp.1315-1318, June 2000.

[13] Mandarin microphone speech corpus – TCC300, http://www.aclclp.org.tw/use_mat.php#tcc300edu.

[14] "HTK Web-Site", http://htk.eng.cam.ac.uk. Accessed 2009.

[15] Bahl, L.R., Brown, R. F., de Souza, P. V., and Mercer, R.L., "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in Proc. ICASSP 1986, pp. 49-52.

[16] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W., Tokuda, K., The HMM-based speech synthesis system version 2.0, Proc. of ISCA SSW6, Bonn, Germany, Aug. 2007.