ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken
Language Processing (ISCSLP 2008)
Kunming, China
December 16-19, 2008

SIMULTANEOUS ACOUSTIC, PROSODIC, AND PHRASING MODEL TRAINING FOR TTS CONVERSION SYSTEMS

Keiichiro Oura[1,3], *Yoshihiko Nankaku*[1], *Tomoki Toda*[2,3], *Keiichi Tokuda*[1,3],
Rannierry Maia[3], *Shinsuke Sakai*[3], *Satoshi Nakamura*[3]
[1]Nagoya Institute of Technology Department of Computer Science and Enginnering,
Gokiso-cho, Showa-ku, Nagoya
[2]Nara Institute of Science and Technology,
8916-5 Takayama-cho, Ikoma, Nara
[3]NiCT/ATR Spoken Language Communication Research Laboratories,
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto

ABSTRACT

A new integrated model for simultaneous modeling of linguistic and acoustic models, and a training algorithm is proposed. Usually, text-to-speech (TTS) systems based on the hidden Markov model (HMM) consist of text analysis and speech synthesis modules. Linguistic and acoustic model training are performed independently using different training data sets. Integrated model parameters were simultaneously optimized by the proposed training algorithm. The derived algorithm optimizes two model parameter sets simultaneously. Therefore, the appropriate model is estimated because we can directly-formulate the TTS problem in which the speech waveform is generated from a word sequence. We conducted objective evaluation experiments using phrasing and prosodic models to evaluate the effectiveness of the proposed technique.

Index Terms— TTS system, hidden Markov model, phrasing model, prosodic model

1. INTRODUCTION

Standard text-to-speech (TTS) systems consist of two major modules: text analysis and speech synthesis modules. Conventionally, these two modules are constructed independently. The text analysis module is trained using text corpora. The module includes phrasing and prosodic models. On the other hand, the speech synthesis module is trained using a labeled speech database. The module includes acoustic models used for speech synthesis, which are based on the hidden Markov model (HMM). Therefore, if these two modules were combined and trained simultaneously as a unified model, we would expect improved overall performance of a TTS system.

In this paper, we define a new integrated model for simultaneous linguistic and acoustic modeling. Two model parameter sets were simultaneously optimized by the proposed training algorithm. In this manner, we directly-formulate the TTS problem of synthesizing a speech waveform from a word sequence. Another advantage of the proposed approach is that hand-labeling of phrasing and prosodic events not required for neither linguistic nor acoustic model training because these labels are regarded as latent variables in the model.

The remainder of this paper is organized as follows. The linguistic model assumed in this paper is described in Section 2. The theoretical framework for integrating linguistic and acoustic models is described in Section 3. An algorithm for training the integrated model is shown in Section 4. Objective evaluation results are shown in Section 5. Finally, concluding remarks and future plans are presented in Section 6.

2. LINGUISTIC MODEL

Text analysis modules in TTS systems consist of several parts (e.g., pronunciation, part-of-speech (POS) tagging, phrasing, and prosodic models), and we call the set of those parts a "linguistic model." In this study, phrasing and prosodic models in particular are used as the "linguistic model" in accordance with tones and break indices "ToBI" [1].

We used two phrasing models. The first model is based on a 7-gram model, and the second model is based on a 4-gram POS model. Two types of pitch events are marked by the prosodic model: pitch events associated with accented syllables (pitch accents) and pitch events associated with intonational boundaries (phrasal tones). Therefore, we used two decision trees for the prosodic model in this paper.

3. INTEGRATION OF LINGUISTIC AND ACOUSTIC MODELS

In this section, we define a new integrated model to optimize linguistic and acoustic models simultaneously. First, a linguistic and an acoustic model are defined. The likelihood of the linguistic model λ_W, e.g., N-gram, decision tree model, is written as $P(\boldsymbol{L} \mid \boldsymbol{W}, \lambda_W)$, where $\boldsymbol{L}$ and $\boldsymbol{W}$ are label sequence and word sequence, respectively. On the other hand, the likelihood of the acoustic model λ_H is given by

$$P(\boldsymbol{O} \mid \boldsymbol{L}, \lambda_H) = \sum_{\boldsymbol{q}} P(\boldsymbol{O} \mid \boldsymbol{q}, \lambda_H) P(\boldsymbol{q} \mid \boldsymbol{L}, \lambda_H), \quad (1)$$

PROOF

where $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T)$ and $\boldsymbol{q} = (q_1, q_2, \ldots, q_T)$ are observation vector sequence and state sequence, respectively.

An integrated model λ that directly models the observation vector sequence $\boldsymbol{O}$ for the word sequence $\boldsymbol{W}$ is derived by combining the linguistic model λ_W and acoustic model λ_H, as follows:

$$P(\boldsymbol{O} \mid \boldsymbol{W}, \lambda) = \sum_{\boldsymbol{L}} \sum_{\boldsymbol{q}} P(\boldsymbol{O} \mid \boldsymbol{q}, \lambda_H) P(\boldsymbol{q} \mid \boldsymbol{L}, \lambda_H) P(\boldsymbol{L} \mid \boldsymbol{W}, \lambda_W), \quad (2)$$

where

$$\lambda = \{\lambda_H, \lambda_W\}. \quad (3)$$

We performed the linguistic model λ_W and acoustic model λ_H training simultaneously to optimize all parameters of the integrated model λ. The derivation of the algorithm is shown in the next Section. Differences between the conventional and proposed training criteria are shown in Fig. 1 and Fig. 2, respectively. In the conventional model, training data has to be labeled by hand or an automatic labeling tool, which is time consuming or causes labeling errors, respectively. On the other hand, in the proposed model, the label sequence is regarded as a latent variable and marginalized like a state sequence in HMM. Labeling the training data accordingly in not necessary.

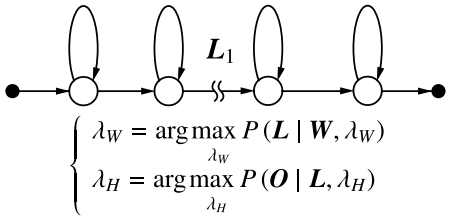

Fig. 1. *Conventional model optimization*

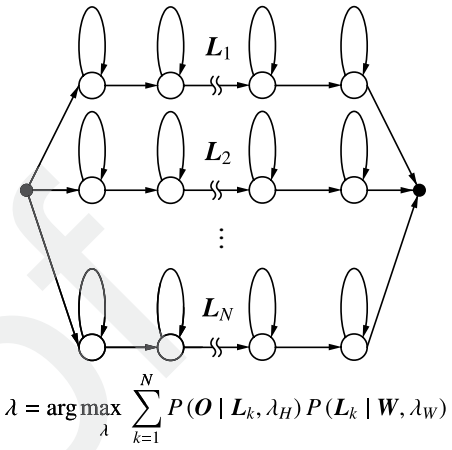

Fig. 2. *Proposed model optimization*

4. PARAMETER ESTIMATION FORMULAS

4.1. EM algorithm

The expectation maximization (EM) algorithm [2] was used for training the proposed model. In the EM algorithm, the likelihood is maximized at each iteration using an auxiliary function called the Q-function:

$$Q(\lambda, \lambda') = \sum_{\boldsymbol{L}} \sum_{\boldsymbol{q}} P(\boldsymbol{q}, \boldsymbol{L} \mid \boldsymbol{O}, \boldsymbol{W}, \lambda_H, \lambda_W) \log\left[P(\boldsymbol{O} \mid \boldsymbol{q}, \lambda_H') P(\boldsymbol{q} \mid \boldsymbol{L}, \lambda_H') P(\boldsymbol{L} \mid \boldsymbol{W}, \lambda_W')\right], \quad (4)$$

where λ, λ', and $P(\boldsymbol{q}, \boldsymbol{L} \mid \boldsymbol{O}, \boldsymbol{W}, \lambda_H, \lambda_W)$ are the integrated model before updating, that after updating, and posterior probabilities of state sequence $\boldsymbol{q}$ and label $\boldsymbol{L}$, respectively. Posterior probabilities are calculated by Bayes' rule:

$$P(\boldsymbol{q}, \boldsymbol{L} \mid \boldsymbol{O}, \boldsymbol{W}, \lambda) = \frac{P(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{L} \mid \boldsymbol{W}, \lambda)}{\sum_{\boldsymbol{L}} \sum_{\boldsymbol{q}} P(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{L} \mid \boldsymbol{W}, \lambda)} = \frac{P(\boldsymbol{O} \mid \boldsymbol{q}, \lambda_H) P(\boldsymbol{q} \mid \boldsymbol{L}, \lambda_H) P(\boldsymbol{L} \mid \boldsymbol{W}, \lambda_W)}{\sum_{\boldsymbol{L}} \sum_{\boldsymbol{q}} P(\boldsymbol{O} \mid \boldsymbol{q}, \lambda_H) P(\boldsymbol{q} \mid \boldsymbol{L}, \lambda_H) P(\boldsymbol{L} \mid \boldsymbol{W}, \lambda_W)}. \quad (5)$$

Increasing the value of the Q-function causes an increase in the likelihood of the training data:

$$Q(\lambda, \lambda') \geq Q(\lambda, \lambda) \Rightarrow P(\boldsymbol{O} \mid \lambda') \geq P(\boldsymbol{O} \mid \lambda). \quad (6)$$

Hence, maximization of the Q-function value at each iteration maximizes the likelihood of the training data. The EM algorithm starts with an initial model parameter λ^0, and iterates between the following two steps:

E-step : compute $Q(\lambda, \lambda^{(t)})$
M-step : $\lambda^{(t+1)} = \arg\max_{\lambda} Q(\lambda, \lambda^{(t)})$

where t denotes the number of the iteration. In this procedure, each step increases the value of the Q-function. Therefore, the likelihood of the training data either increases or remains unchanged at each iteration.

In the M-step of the integrated model λ, the linguistic model λ_W and the acoustic model λ_H were updated individually. The Q-function of the linguistic model λ_W is defined as

$$Q_W(\lambda, \lambda') = \sum_{\boldsymbol{L}} P(\boldsymbol{L} \mid \boldsymbol{O}, \boldsymbol{W}, \lambda) \log P(\boldsymbol{L} \mid \boldsymbol{W}, \lambda_W'). \quad (7)$$

Proof

The acoustic model λ_H was optimized by setting derivatives of the Q-function to zero. As a result, the mean $\boldsymbol{\mu}_i$ and variance $\boldsymbol{\Sigma}_i$ of the i-th state output probability distribution (Gaussian distribution) were estimated as:

$$\boldsymbol{\mu}_i = \frac{\sum_t \sum_{\boldsymbol{L}} \gamma_i(t, \boldsymbol{L})\, \boldsymbol{o}_t}{\sum_t \sum_{\boldsymbol{L}} \gamma_i(t, \boldsymbol{L})} \quad (8)$$

and

$$\boldsymbol{\Sigma}_i = \frac{\sum_t \sum_{\boldsymbol{L}} \gamma_i(t, \boldsymbol{L})\, (\boldsymbol{o}_t - \boldsymbol{\mu}_i)(\boldsymbol{o}_t - \boldsymbol{\mu}_i)^\top}{\sum_t \sum_{\boldsymbol{L}} \gamma_i(t, \boldsymbol{L})}, \quad (9)$$

respectively, where

$$\begin{aligned} \gamma_i(t, \boldsymbol{L}) &= P(q_t = i, \boldsymbol{L} \mid \boldsymbol{O}, \boldsymbol{W}, \lambda) \\ &= P(q_t = i \mid \boldsymbol{L}, \boldsymbol{O}, \lambda)\, P(\boldsymbol{L} \mid \boldsymbol{O}, \boldsymbol{W}, \lambda). \end{aligned} \quad (10)$$

Posterior probabilities $P(q_t = i \mid \boldsymbol{L}, \boldsymbol{O}, \boldsymbol{W}, \lambda)$ of state q_t were computed by the forward-backward algorithm [3] using label sequence $\boldsymbol{L}$. On the other hand, posterior probabilities $P(\boldsymbol{L} \mid \boldsymbol{O}, \boldsymbol{W}, \lambda)$ of label sequence $\boldsymbol{L}$ were written as follows:

$$P(\boldsymbol{L} \mid \boldsymbol{O}, \boldsymbol{W}, \lambda) = \frac{P(\boldsymbol{L} \mid \boldsymbol{W}, \lambda)\, P(\boldsymbol{O} \mid \boldsymbol{L}, \lambda)}{\sum_{\boldsymbol{L}} P(\boldsymbol{L} \mid \boldsymbol{W}, \lambda)\, P(\boldsymbol{O} \mid \boldsymbol{L}, \lambda)}. \quad (11)$$

4.2. N-best approximation

Direct implementation of the EM algorithm is not feasible because the total number of possible combinations of label sequence $\boldsymbol{L}$ is too large. Thus, the N-best hypotheses generated by the text analysis module were used in this study[1]. The E-step was implemented accordingly as follows:

1 : generate N-best label sequences $\boldsymbol{L}_i, i = 1, \ldots, N$
2 : compute $P(\boldsymbol{O} \mid \boldsymbol{L}_i, \boldsymbol{W}, \lambda_H)$ for each label sequences $\boldsymbol{L}_i$
3 : compute $P(\boldsymbol{L}_i \mid \boldsymbol{O}, \boldsymbol{W}, \lambda)$ for each label sequences $\boldsymbol{L}_i$
4 : compute $Q\left(\lambda, \lambda'\right)$

In the M-step, model parameters were updated using the N-best label sequences. The above procedure optimizes the linguistic and acoustic models simultaneously. Furthermore, a state-sharing structure of HMM that matches the linguistic model was constructed by a context-clustering technique [4].

[1] Although variational approximation is one of the methods for solving this problem, we chose the N-best approximation because label sequence $\boldsymbol{L}$ is strongly correlated with state sequence $\boldsymbol{q}$.

5. EXPERIMENT

5.1. Experimental conditions

Objective evaluations were conducted on the CMU-ARCTIC speech database to evaluate the performance of the proposed system. Training data, testing data, and speech analysis conditions are shown in Table 1. Each feature vector consisted of spectrum and F_0 parameter vectors. Each spectrum parameter vector consisted of the 0th - 39th STRAIGHT [5] mel-cepstral coefficients, their delta coefficients, and delta-delta coefficients. The F_0 parameter vector consisted of log F_0, its delta coefficient, and delta-delta coefficient. We used a 5-state left-to-right HMM structure with no-skip. Forty-one phonemes including the pause were used as speech units. Context-clustering based on a decision tree was applied to spectrum, F_0, and state duration models, individually. The minimum description length (MDL) criterion [4] was used to stop tree growth.

We trained linguistic models using the Boston University Radio Speech Corpus for the conventional automatic labeling technique. In the proposed system, these models were used as initial linguistic models.

5.2. Evaluation

We calculated the root mean square error (RMSE) and correlation coefficient (Corr) of F_0 contour generation with respect to original speech in the voiced portions of the data. The RMSE and Corr are widely used to measure the accuracy of F_0 contour generation [6, 7, 8, 9][2]. In this experiment, 4 systems were constructed as follows:

BASELINE: Only acoustic models were trained. The one-best label sequence was used.

PROSODIC: Both acoustic and prosodic models were trained simultaneously.

PHRASING: Both acoustic and phrasing models were trained simultaneously.

PROSODIC + PHRASING: Acoustic, prosodic, and phrasing models were trained simultaneously.

Table 1. *Experimental conditions*

Database	CMU-ARCTIC speech database a female speaker SLT 1132 sentences train : 1000 sentences test : 132 sentences
Sampling rate	16kHz
Frame shift	5ms
Window length	25ms
Window function	Blackman window

[2] Although subjective evaluation experiments are also required, they have to be postponed because finding a sufficient number of native English speakers was not easy.

Proof

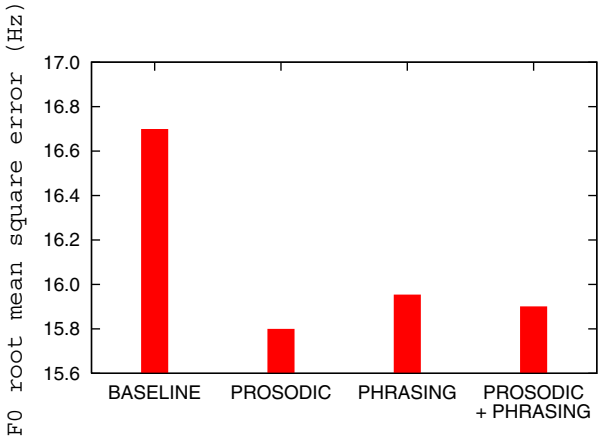

Fig. 3. *F_0 RMSE results*

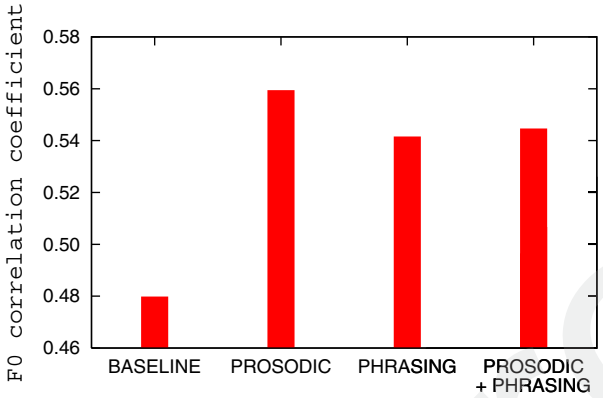

Fig. 4. *F_0 Corr results*

In the *N*-best approximation of simultaneous linguistic and acoustic model training, 100-best label hypotheses were used. About 20 days were taken to train the integrated models of the proposed system.

Calculations of RMSE and Corr are shown in Fig. 3 and Fig. 4, respectively. Three systems, PROSODIC, PHRASING, and PROSODIC + PHRASING, using simultaneous training of linguistic and acoustic models achieved a smaller RMSE and larger Corr, than those of "BASELINE." A graph of F_0 contours is shown in Fig. 5. The F_0 contour generated by "PROSODIC" seems to exhibit a better goodness of fit with respect to original speech than that of "BASELINE."

6. CONCLUSION

In this paper, we defined a new integrated model in which linguistic and acoustic models were combined into one model, and all model parameters were estimated simultaneously by the proposed training algorithm. We conducted objective evaluation experiments using phrasing and prosodic models as linguistic models to evaluate the effectiveness of the

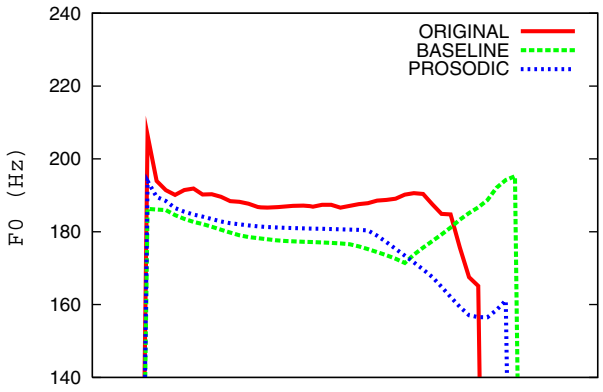

Fig. 5. *F_0 contours*

proposed system. The results demonstrate that the proposed system achieves better F_0 modeling accuracy than that of the conventional system. Future work will include simultaneous training of POS tagging modules and acoustic models. Subjective listening tests performed by native English speakers on a large database are also planned.

7. REFERENCES

[1] K. Silverman, *et al.*, "ToBI: A standard for labeling English prosody," Proc. of ICSLP, pp.867-870, 1992.

[2] A. P. Dempster, *et al.*, "Maximum-likelihood from incomplete data via the EM algorithm," J. Royal Statist. Soc., Ser. B, 39, pp.1-38, 1977.

[3] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. of IEEE, vol.77, no.2 , pp.257-285, 1989.

[4] K. Shinoda, T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," J. Acoust. Soc. Jpn. (E), 21 (2), pp.79-86, 2000.

[5] H. Kawahara, *et al.*, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, 27, pp.187-207, 1999.

[6] A. Black, A. J. Hunt, "Generating F_0 contours from tobi labels using linear regression," Proc. of ICSLP96, pp.1385-1388, 1996.

[7] K. Dusterhoff, A. W. Black, P. Taylor, "Using decision trees within the tilt intonation model to predict F_0 contours," Proc. of EUROSPEECH99, pp.1627-1630, 1999.

[8] X. Sun, "F_0 generation for speech synthesis using a multi-tier approach," Proc. of ICSLP02, pp.2077-2080, 2002.

[9] S. Sakai, "Additive Modeling of English F_0 Contour for Speech Synthesis," Proc. of ICASSP2005, vol.I, pp.277-280, 2005.

Proof