Journal of Computer Science 7 (3): 359-365, 2011
ISSN 1549-3636

Analysis of Decision Trees in Context Clustering of Hidden Markov Model Based Thai Speech Synthesis

Suphattharachai Chomphan
Department of Electrical Engineering, Faculty of Engineering at Si Racha,
Kasetsart University, 199 M.6, Tungsukhla, Si Racha, Chonburi, 20230, Thailand

Abstract: Problem statement: In Thai speech synthesis using Hidden Markov model (HMM) based synthesis system, the tonal speech quality is degraded due to tone distortion. This major problem must be treated appropriately to preserve the tone characteristics of each syllable unit. Since tone brings about the intelligibility of the synthesized speech. It is needed to establish the tone questions and other phonetic questions in tree-based context clustering process accordingly. **Approach:** This study describes the analysis of questions in tree-based context clustering process of an HMM-based speech synthesis system for Thai language. In the system, spectrum, pitch or F0 and state duration are modeled simultaneously in a unified framework of HMM, their parameter distributions are clustered independently by using a decision-tree based context clustering technique. The contextual factors which affect spectrum, pitch and duration, i.e., part of speech, position and number of phones in a syllable, position and number of syllables in a word, position and number of words in a sentence, phone type and tone type, are taken into account for constructing the questions of the decision tree. All in all, thirteen sets of questions are analyzed in comparison. **Results:** In the experiment, we analyzed the decision trees by counting the number of questions in each node coming from those thirteen sets and by calculating the dominance score given to each question as the reciprocal of the distance from the root node to the question node. The highest number and dominance score are of the set of phonetic type, while the second, third highest ones are of the set of part of speech and tone type. **Conclusion:** By counting the number of questions in each node and calculating the dominance score, we can set the priority of each question set. All in all, the analysis results bring about further development of Thai speech synthesis with efficient context clustering process in an HMM-based speech synthesis system.

Key words: Thai speech synthesis, tree-based context clustering, HMM-based speech synthesis, Hidden Markov Model (HMM), Multi-Space probability Distribution (MSD), Minimum Description Length (MDL), synthesis framework

INTRODUCTION

Thai speech synthesis has been developed for years since it is one of the key technologies for realizing natural human-computer interaction for Thai. However the systematic analysis of the decision trees in the context clustering process has not been thoroughly carried out yet. This study could bring about an appropriate construction of question sets for the decision trees. In the other words, this study consequently aims at improving the synthetic speech quality (Chomphan, 2009; 2010a; 2010b).

In the HMM-based speech synthesis framework, the core system consists of a training stage and a synthesis stage (Tokuda *et al*., 1995; Masuko *et al*., 1996; Yoshimura *et al*., 1999). The context dependent HMMs are constructed in the training stage, subsequently the synthesized speech is generated using the sentence HMMs associated with the arbitrary given texts. In the training stage, the context clustering is an important process to treat the problem of limitation of training data. Information sharing of training data in the same cluster or the terminal node (tree leaf) in the decision-tree-based context clustering is the essential concept, therefore construction of contextual factors and design of tree structure for the decision-tree-based context clustering must be done appropriately. This study focuses mainly on the analysis of the decision trees in the context clustering process. The question that appears in all nodes of the decision trees are statistically investigated and then the tree occupancy by the question sets are analyzed comparatively.

MATERIALS AND METHODS

HMM-based speech synthesis: A block-diagram of the HMM-based TTS system is shown in Fig. 1. The system consists of two stages including the training

stage and the synthesis stage (Tamura *et al*., 2001; Yamagishi *et al*., 2003). In the training stage, mel-cepstral coefficients are extracted at each analysis frame as the static features from the speech database. Then the dynamic features, i.e., delta and delta-delta parameters, are calculated from the static features. Spectral parameters and pitch observations are combined into one observation vector frame-by-frame and speaker dependent phoneme HMMs are trained using the observation vectors. To model variations of spectrum, pitch and duration, phonetic and linguistic contextual factors, such as phoneme identity factors, are taken into account (Yoshimura *et al*., 1999). Spectrum and pitch are modeled by multi-stream HMMs and output distributions for spectral and pitch parts are continuous probability distribution and Multi-Space probability Distribution (MSD) (Tokuda *et al*., 1999), respectively. Then, a decision tree based context clustering technique is separately applied to the spectral and pitch parts of context dependent phoneme HMMs (Young *et al*., 1994). Finally state durations are modeled by multi-dimensional Gaussian distributions and the state clustering technique is also applied to the duration distributions (Yoshimura *et al*., 1998).

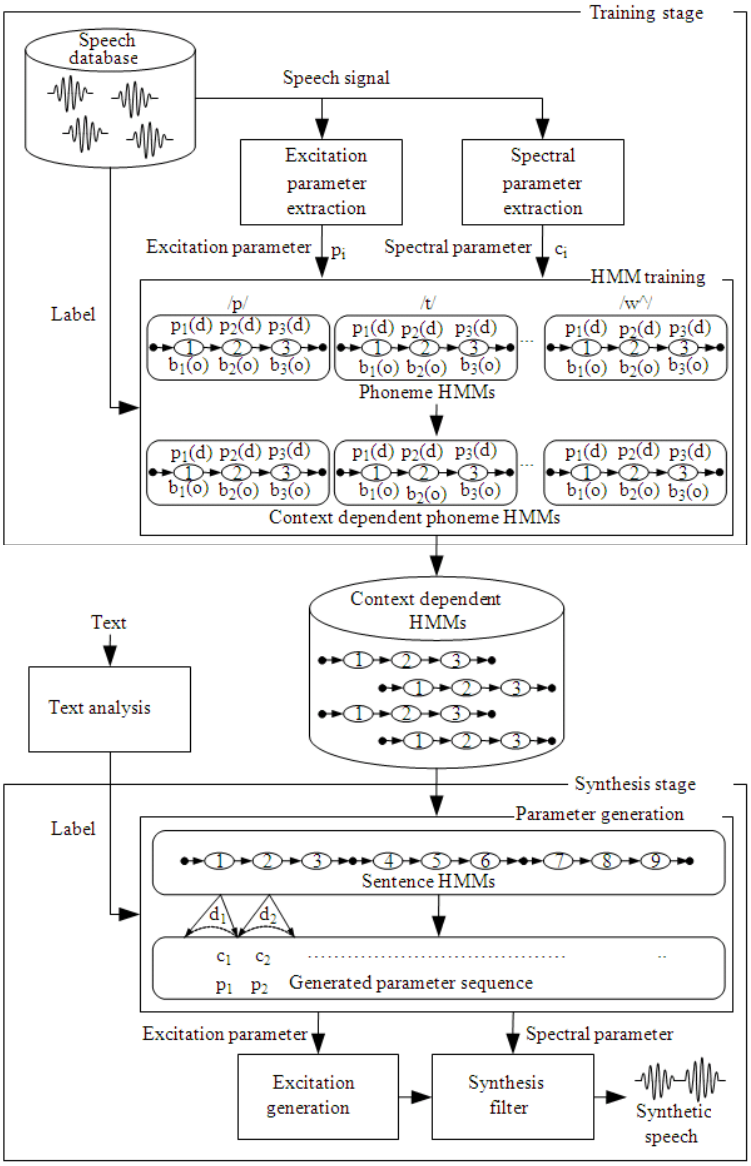

Fig. 1: A block diagram of an HMM-based speech synthesis system

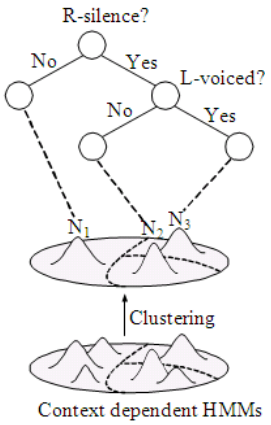

Fig. 2: An example of decision tree

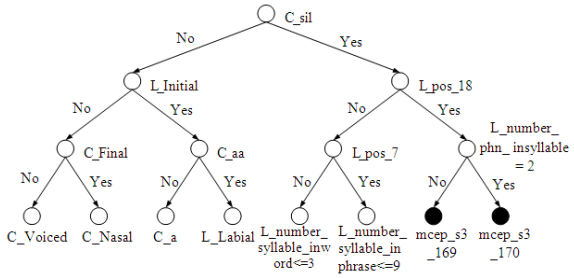

Fig. 3: An example of the constructed decision tree

In the synthesis stage, first, an arbitrary given text to be synthesized is transformed into context dependent phoneme label sequence. According to the label sequence, a sentence HMM, which represents the whole text to be synthesized, is constructed by concatenating adapted phoneme HMMs. From the sentence HMM, phoneme durations are determined based on state duration distributions (Yoshimura *et al*., 1998). Then spectral and pitch parameter sequences are generated using the algorithm for speech parameter generation from HMMs with dynamic features (Tokuda *et al*., 1995; 2000). Finally by using MLSA filter (Fukuda *et al*., 1992; Taleizadeh *et al*., 2009), speech is synthesized from the generated mel-cepstral and pitch parameter sequences.

Decision-tree-based context clustering: In the training stage, context dependent models taking account of several combinations of contextual factors are constructed. However, as the number of contextual factors we take into account increases, their combinations also increase exponentially. Therefore, model parameters with sufficient accuracy cannot be estimated with limited training data. In other words, it is impossible to prepare the speech database which includes all combinations of contextual factors. To overcome this problem, the decision-tree based context clustering technique is employed to the distributions of the associated speech features.

The implemented decision tree is a binary tree, where a question splitting contexts into two sub-groups is prepared within each of an intermediate node and the Minimum Description Length (MDL) criterion is used for selecting nodes to be split. All contexts can be found by traversing the tree, starting from the root node then selecting the next node depending on the answer to a question about the current context. Therefore, if once the decision tree is constructed, unseen contexts can be prepared (Riley, 1989; Yoshimura *et al*., 1998).

In continuous speech context, parameter sequences of particular speech unit (e.g., phoneme) vary depending on phonetic context. To treat the variations appropriately, context dependent models, such as triphone models, are usually employed. In the HMM-based speech synthesis system, we use speech units considering prosodic and linguistic context such as syllable, phrase, part of speech and sentence information to model suprasegmental features in prosodic feature appropriately. However, it is impossible to prepare training data which cover all possible context dependent units, moreover, there is great variation in the frequency of appearance of each context dependent unit. To alleviate these problems, several techniques are proposed to cluster HMM states and share model parameters among states in each cluster. In this study, we exploit a decision-tree-based state tying algorithm. This algorithm is referred to as the decision-tree-based context clustering algorithm.

The binary decision tree is described as follows. An example of the decision tree is shown in Fig. 2. Each non-terminating node has a context related question, such as R-silence? (“is the succeeding phoneme a silence?”) or L-voiced? (“is the preceding phoneme a voiced phoneme?”) and two child nodes representing “yes” and “no” answers to the question. All leaf nodes have state output distributions. Using the decision-tree-based context clustering, model parameters of the speech units for any unseen contexts can be obtained, because any context can reach one of the leaf nodes by going down the tree starting from the root node then selecting the next node depending on the answer about the current context. An example of the decision tree for the spectral part constructed in the context clustering process for Thai is shown in Fig. 3.

Constructed contextual factors: Contextual information is language dependent. Besides, a large number of contextual factors do not guarantee the synthesized speech with better quality. There are several contextual factors that affect spectrum, F0 pattern and duration, e.g., phone identity factors, locational factors (Yogameena *et al*., 2010). There should be efficient factors for a certain language to model context dependent HMMs. Thirteen contextual factor sets in 5 levels of speech unit were constructed according to 2 sources of information, including the phonological information (for phoneme and syllable levels) and the utterance structure from Thai text corpus named ORCHID (for word, phrase and utterance levels).

- Phoneme level
  S1. {preceding, current, succeeding} phonetic type
  S2. {preceding, current, succeeding} part of syllable structure
- Syllable level
  S3. {preceding, current, succeeding} tone type
  S4. The number of phones in {preceding, current, succeeding} syllable
  S5. Current phone position in current syllable
- Word level
  S6. Current syllable position in current word
  S7. Part of speech
  S8. The number of syllables in {preceding, current, succeeding} word
- Phrase level
  S9. Current word position in current phrase
  S10. The number of syllables in {preceding, current, succeeding} phrase
- Utterance level
  S11. Current phrase position in current sentence
  S12. The number of syllables in current sentence
  S13. The number of words in current sentence

Subsequently, these contextual information sets were transformed into question sets which finally applied at the context clustering process in the training stage with the total question number of 1156.

RESULTS

Experimental conditions: A set of phonetically balanced sentences of Thai speech database named TSynC-1 from NECTEC (Hansakunbuntheung *et al*., 2005; Subramanian *et al*., 2010) was used for training HMMs. The whole sentence text was collected from Thai part-of-speech tagged ORCHID corpus. The speech in the database was uttered by a professional female speaker with clear articulation and standard Thai accent. The phoneme labels included in TSynC-1 and the utterance structure from ORCHID were used to construct the context dependent labels with 79 different phonemes including silence and pause in the case of tone-independent phonemes and 246 different phonemes including silence and pause in the case of tone-dependent phonemes. Another male speech set has been used with the same condition.

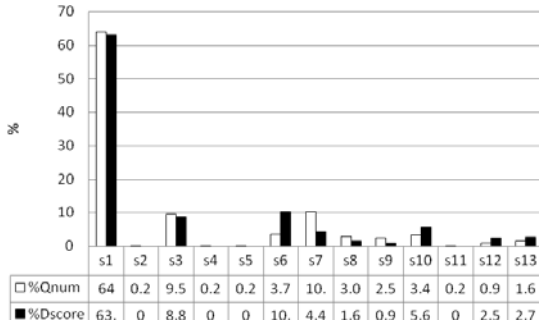

Fig. 4: Question Numbers (Qnum in %) and Dominance scores (Dscore in %) in state duration tree for female speech database

Speech signal were sampled at a rate of 16kHz and windowed by a 25ms Blackman window with a 5ms shift. Then mel-cepstral coefficients were extracted by mel-cepstral analysis. The feature vectors consisted of 25 mel-cepstral coefficients including the zeroth coefficient, logarithm of F0 and their delta and delta-delta coefficients (Tokuda *et al*., 1995; Alfred, 2009).

We used 5-state left-to-right Hidden Semi-Markov Models (HSMMs) in which the spectral part of the state was modeled by a single diagonal Gaussian output distribution (Zen *et al*., 2004). Using the HSMMs, the explicit state duration probability is incorporated into HMMs and the state duration probability is reestimated by using EM algorithm (Russell and Moore, 1985). Note that each context dependent HSMM corresponds to a phoneme-sized speech unit. The numbers of training utterances for both speech sets are 500.

To analyze the contribution of each set of contextual factors, we explored 3 decision trees generated in the clustering process at the training stage of the system including spectrum, F0 and state duration trees. Two criteria were taken into account. First, the number of the existing questions in each set was counted. The 3 highest proportions among 13 sets are shown in Fig. 4-6. Second, based on the assumption that the question existing near the root node is more important than the further one, a dominance score given to each question was calculated as the reciprocal of the distance from the root node to the question node. Subsequently, the dominance scores for each question set were summed up. The 3 highest proportions among 13 sets are shown in Fig. 7-9.

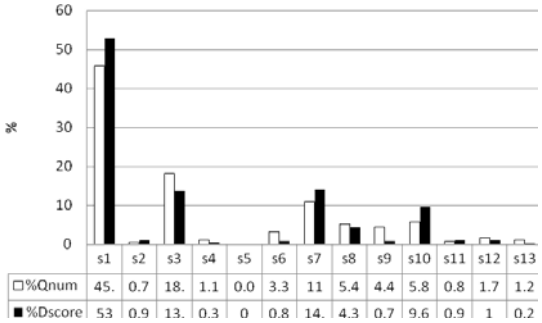

Fig. 5: Question Numbers (Qnum in %) and Dominance scores (Dscore in %) in F0 tree for female speech database

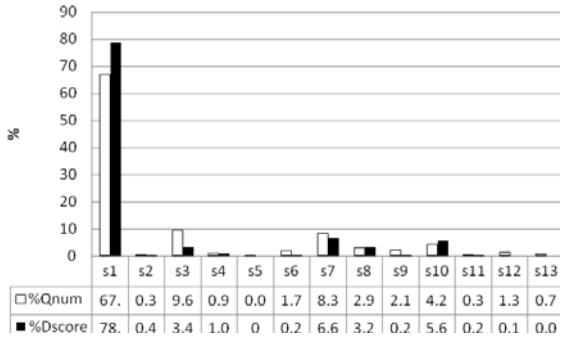

Fig. 6: Question Numbers (Qnum in %) and Dominance scores (Dscore in %) in spectrum tree for female speech database

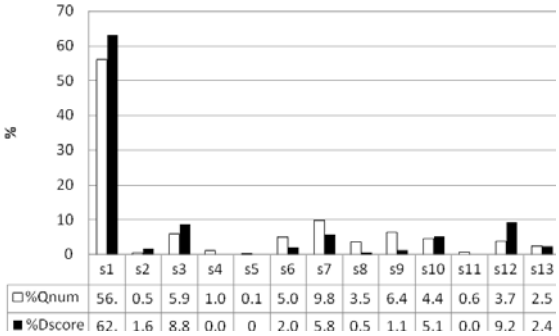

Fig. 7: Question Numbers (Qnum in %) and Dominance scores (Dscore in %) in state duration tree for male speech database

Table 1: Question Numbers (Qnum in quantity) and Dominance scores (Dscore in points) in all tree types

Trees	Qnum in total	Dscore in total
Female state duration	430	12.72
Female F0	2408	62.70
Female spectrum	2219	59.45
Male state duration	590	11.81
Male F0	4318	72.34
Male spectrum	3918	65.27

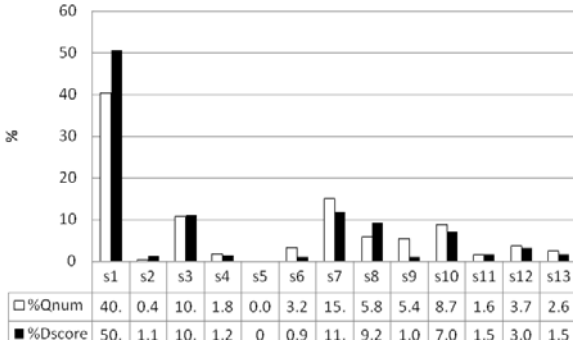

Fig. 8: Question Numbers (Qnum in %) and Dominance scores (Dscore in %) in F0 tree for male speech database

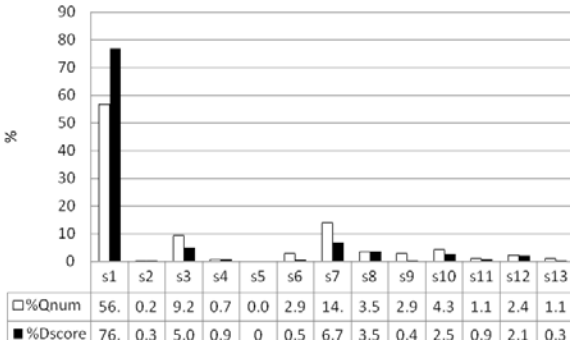

Fig. 9: Question Numbers (Qnum in %) and Dominance scores (Dscore in %) in spectrum tree for male speech database

Question numbers and dominance scores for all tree types are consequently summarized in Table 1. The highest question number belongs to male F0 tree, while lowest question number belongs to female state duration tree. As for the dominance score, the highest dominance score belongs to male F0 tree, while lowest dominance score belongs to male state duration tree.

DISCUSSION

It can be seen from Fig. 4-9 that some question sets have higher proportions of tree occupancy, but have lower proportions of dominance and vice versa. In decision tree construction, we adopted a top-down sequential optimization procedure (Young *et al*., 1994; Ping *et al*., 2009), where the question that gives the best split of the current node (i.e., gives the maximum increase in log likelihood) is selected. This leaded to the second criterion where the reciprocal of the distance from the root node to the question node was used as a weighting factor. On the other hand, the first criterion

used a constant value as a weighting factor. From this context, the second criterion is supposed to be more meaningful than the first one. However, there is no explicit study to indicate which criterion is the most appropriate for tree analysis.

Considering the first criteria from Fig. 4-9, it can be seen that the most tree-occupied question sets are phonetic type (S1), tone type (S3) and part of speech (S7), for nearly all trees. Male and female have the same results with a little difference of the order of the second and third places. As for the second criteria from Fig. 4-6 for female speech, the most dominant question sets are little different for each decision trees, i.e., phonetic type (S1), part of speech (S7) and the number of syllables in {preceding, current, succeeding} phrase (S10) for spectrum tree, phonetic type (S1), part of speech (S7) and tone type (S3) for F0 tree and phonetic type (S1), the current syllable position in current word (S6) and tone type (S3) for state duration tree. As for the second criteria from Fig. 7-9 for male speech, the most dominant question sets are little different for each decision trees, i.e., phonetic type (S1), part of speech (S7) and tone type (S3) for spectrum and F0 trees and phonetic type (S1), the number of syllables in current sentence (S12) and tone type (S3) for state duration tree. When considering the contribution of tone type question set (S3), the F0 tree is affected most among all 3 trees.

CONCLUSION

In this study, we describe the analysis of questions in tree-based context clustering process of an HMM-based speech synthesis system for Thai language. In the system, spectrum, pitch and state duration are modeled simultaneously in a unified framework of HMM. The contextual factors which affect spectrum, pitch and duration, i.e., part of speech, position and number of phones in a syllable, position and number of syllables in a word, position and number of words in a sentence, phone type and tone type, are taken into account for constructing the questions of the decision tree. In the experiment, the highest number and dominance score are of the set of phonetic type, while the second, third highest ones are of the set of part of speech and tone type mostly. All in all, by counting the number of questions in each node and calculating the dominance score, we can set the priority of each question set. The analysis results bring about further development of Thai speech synthesis with efficient context clustering process in an HMM-based speech synthesis system.

ACKNOWLEDGEMENT

The researcher is grateful to NECTEC for providing the TSynC-1 speech database.

REFERENCES

Alfred, R., 2009. Optimizing feature construction process for dynamic aggregation of relational attributes. J. Comput. Sci., 5: 864-877. DOI: 10.3844/jcssp.2009.864.877

Chomphan, S., 2009. Towards the development of speaker-dependent and speaker-independent hidden markov model-based Thai speech synthesis. J. Comput. Sci., 5: 905-914. DOI: 10.3844/jcssp.2009.905.914

Chomphan, S., 2010a. Tone question of tree based context clustering for hidden Markov model based Thai speech synthesis. J. Comput. Sci., 6: 1468-1472. DOI: 10.3844/jcssp.2010.1468.1472

Chomphan, S., 2010b. Performance evaluation of multi-pulse based code excited linear predictive speech coder with bitrate scalable tool over additive white gaussian noise and rayleigh fading channels. J. Comput. Sci., 6: 1438-1442. DOI: 10.3844/jcssp.2010.1433.1437

Fukuda, T., K. Tokuda, T. Kobayashi and S. Imai, 1992. An adaptive algorithm for mel-cepstral analysis of speech. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 23-26, San Francisco, CA, USA., pp: 137-140.

Hansakunbuntheung, C., A. Rugchatjaroen and C. Wutiwiwatchai, 2005. Space reduction of speech corpus based on quality perception for unit selection speech synthesis. Proceedings of the International Symposium on Natural Language Processing, Dec. 2005, Bangkok, Thailand, pp: 127-132.

Masuko, T., K. Tokuda, T. Kobayashi and S. Imai, 1996. Speech synthesis using HMMs with dynamic features. Proceedings of the 1996 IEEE International Conference on Acoustics, Speech and Signal Processing, May 7-10, Atlanta, GA, USA., pp: 389-392. DOI: 10.1109/ICASSP.1996.541114

Ping, Z., T. Li-Zhen and X. Dong-Feng, 2009. Speech recognition algorithm of parallel subband HMM based on wavelet analysis and neural network. Inform. Technol. J., 8: 796-800

Riley, M.D., 1989. Statistical tree-based modeling of phonetic segment durations. J. Acoust. Soc. Am. 85: 44-44. DOI:10.1121/1.2026979

Russell, M.J. and R.K. Moore, 1985. Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP'85), Worcs, UK., pp: 5-8. DOI: 10.1109/ICASSP.1985.1168477

Subramanian, R., S.N. Sivanandam and C. Vimalarani, 2010. An optimization of design for s4-duty induction motor using constraints normalization based violation technique. J. Comput. Sci., 6: 107-111. DOI: 10.3844/jcssp.2010.107.111

Taleizadeh, A.A., S.T.A. Niaki and M.B. Aryanezhad, 2009. Multi-product multi-constraint inventory control systems with stochastic replenishment and discount under fuzzy purchasing price and holding costs. Am. J. Applied Sci., 6: 1-12. DOI: 10.3844/ajassp.2009.1.12

Tamura, M., T. Masuko, K. Tokuda and T. Kobayashi, 2001. Text-to-speech synthesis with arbitrary speaker's voice from average voice. Proceedings of 7th European Conference on Speech Communication and Technology, (ECSCT'01), Aalborg, Denmark, pp: 345-348.

Tokuda, K., T. Kobayashi and S. Imai, 1995. Speech parameter generation from HMM using dynamic features. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 9-12, Detroit, MI., pp: 660-663. DOI: 10.1109/ICASSP.1995.479684

Tokuda, K., T. Masuko, N. Miyazaki and T. Kobayashi, 1999. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Mar. 15-19, Phoenix, AZ, USA., pp: 229-232. DOI: 10.1109/ICASSP.1999.758104

Tokuda, K., T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, 2000. Speech parameter generation algorithms for HMM-based speech synthesis. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASP'00), Istanbul, Turkey, pp: 1315-1318.

Yamagishi, J., T. Masuko, K. Tokuda and T. Kobayashi, 2003. A training method for average voice model based on shared decision tree context clustering and speaker adaptive training. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Apr. 6-10, Hong Kong, China, pp: 716-719. DOI: 10.1109/ICASSP.2003.1198881

Yogameena, B., S.V. Lakshmi, M. Archana and S.R. Abhaikumar, 2010. Human behavior classification using multi-class relevance vector machine. J. Comput. Sci., 6: 1021-1026. DOI: 10.3844/jcssp.2010.1021.1026

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. IEICE Trans. Inform. Syst., 83: 2099-2107.

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, 1998. Duration modeling for HMM-based speech synthesis. Proceedings of the International Conference on Spoken Language Processing, (ICSLP'98), Sydney, Australia, pp: 29-32.

Young, S. J., J. Odell and P. Woodland, 1994. Tree-based state tying for high accuracy acoustic modelling. Proceedings of the Workshop on Human Language Technology, (WHLT'94), Association for Computational Linguistics Stroudsburg, PA, USA., pp: 307-312. DOI: 10.3115/1075812.1075885

Zen, H., K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, 2004. Hidden semi-Markov model based speech synthesis. Proceeding of the 8th International Conference on Spoken Language Processing, Oct. 4-8, Jeju Island, Korea, pp: 1393-1396.