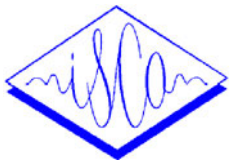

7th ISCA Workshop on
Speech Synthesis (SSW-7)
Kyoto, Japan
September 22-24, 2010

An Investigation of the Impact of Speech Transcript Errors on HMM Voices

Jinfu Ni and Hisashi Kawai

Spoken Language Communication Group, MASTAR project,
National Institute of Information and Communications Technology, Japan
jinfu.ni@nict.go.jp, hisashi.kawai@nict.go.jp

Abstract

Toward automatic creation of web-based voice fonts at low cost, automatic speech transcription technology is used to obtain the linguistic features for building HMM-based voices from audio web contents. This paper presents an investigation of the influences of erroneous transcripts on such voices. We simulate varied speech transcript errors by using a large vocabulary automatic speech recognizer (LVASR) to dictate thousands of Japanese utterances from two speakers (a male and a female). A set of experiments is conducted on dozens of HMM voices built upon both dictated and correct transcripts. The results indicate a significant impact of speech transcript errors on the voices. One direct impact is increasing the number of leaf nodes of the decision trees associated with both state duration and F0 but decreasing that with cepstrum in comparison with the reference voices by correct transcripts. The HMM voice quality in mean opinion scores (MOS) is closely related to the word and phone accuracy of speech transcriptions. To achieve fair voice quality with limited training samples, for example, the word and phone accuracy must be higher than 50% and 80%, respectively.
Index Terms: HMM-based speech synthesis, web-based voice-fonts, unsupervised approach, HTS

1. Introduction

Synthetic speech is a core interface of speech-to-speech (STS) translation systems [1]. There is a growing need for letting users select their favorite voices for personalized STS communication. The HMM (hidden Markov model)-based speech synthesis system (HTS) [2][3] enables rapid building of HMM voices when necessary training data including speech samples and full-context labels representing the underlying linguistic information are provided. Conventionally, to build an HMM voice, a target speaker records her/his speech by reading selected text in a soundproof room. However, when there are large numbers of target speakers —- 10,000, for instance —- the cost of recording speech essentially renders the recording impossible. On the other hand, the continually decreasing cost of storage capacity and increasing access to the internet are aiding the amassing of large volumes of audio contents, including web radio and spoken documents. We have recently been studying methodology for using audio web contents to produce low-cost and varied HMM voices [4] rather than directly recording a large-scale speech corpus. It should be noted that the copyrighting and privacy of audio web contents are still open issues.

Toward automatic creation of web-based HMM voices, automatic speech transcription technology is necessary for obtaining the text and phone transcripts of audio web contents that are used to extract morpheme boundaries, parts-of-speech, and accent types for generating full-context labels. The training of HMM voices basically assumes that the input linguistic features involved in the full-context labels can match the acoustic features of the training samples. When extracting the linguistic features from the automatic speech transcriptions, certain deviation from this assumption cannot be avoided. This paper reports on an experimental investigation of the influences of erroneous transcripts on HMM voices. The rest of the paper is organized as follows. Section 2 outlines a testbed used in this paper. Section 3 describes the experiment method. Sections 4 and 5 respectively present the experiment results and discussions. Section 6 concludes the paper.

2. Testbed outline

Figure 1 shows a diagram of an unsupervised approach with automatic speech transcription technology for creation of web-based HMM voices [4]. This consists of four main components: (1) amassing audio web contents from the internet; (2) extracting unusable speech from them and discarding, if possible, noise, music, and the unusable speech; (3) transcribing the web-based speech for obtaining the underlying linguistic and phone-time alignment information, and (4) creating HMM voices using the speech and the linguistic information. In this paper, we basically use 3 and 4 as a platform to perform the experimental investigation.

2.1. Automatic speech transcription

We use in-house large vocabulary automatic speech recognition (LVASR) (for travel conversation) to transcribe input speech into surface text in a speaker-independent manner and perform phone-time alignment. In this paper, we alter the beam search width of the LVASR to obtain varied word errors, thus simulating transcript errors for the experimental investigation.

2.2. Creation of HMM voices

To build an HMM voice by HTS, the speech samples and their monophone and full-context labels must be provided. Thus the creation of HMM voices consists of the following two steps.
Step 1: Generate monophone and full-context labels.

- Use the phone-time alignment information to generate the monophone labels. The phone transcripts are more accurate than those converted from the surface text.
- Use the front-end of a text-to-speech system [5] to morphologically analyze the surface text and estimate accent types. If some of these accent types fail to be estimated due to word errors, we further estimate them based on the phone transcripts before taking a default accent type.
- Generate full-context labels for the sets of utterances in terms of the morphological analysis results, phone transcripts, and detected pauses.

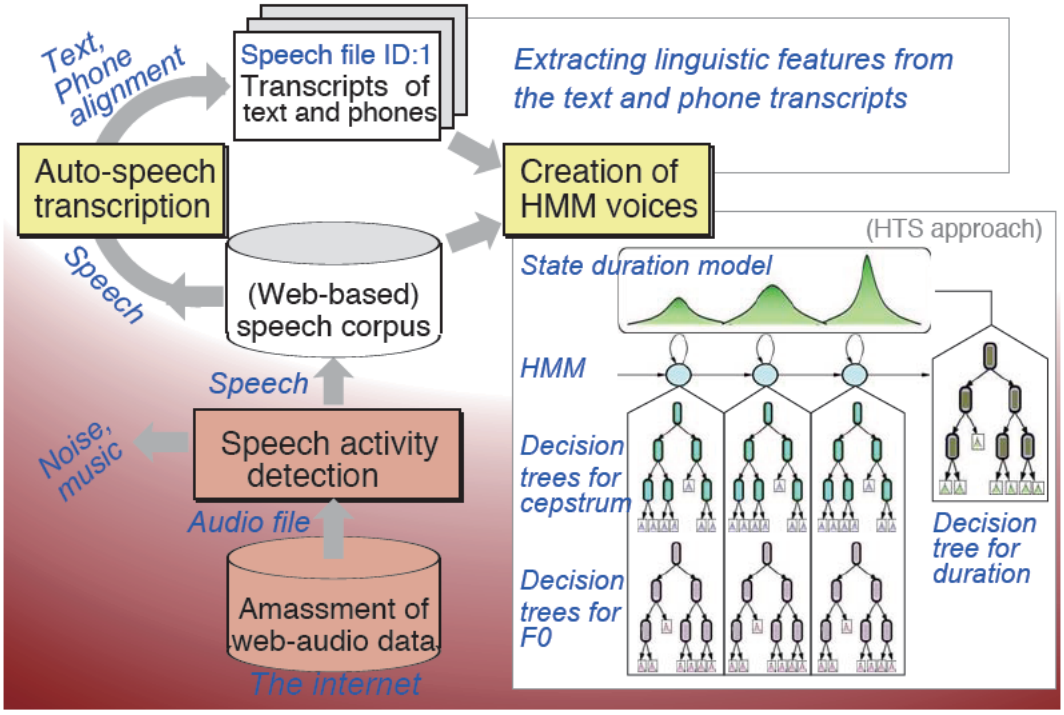

Figure 1: *Testbed with automatic speech transcription technology for building HMM voices using audio web contents.*

Step 2: Train HMM voices using HTS [3] as usual.

- Use the default parameters available in the HTS-2.1 demo scripts including GV [6].
- Use 24-order mel-cepstrum analyzed at 5 ms frames.
- Use the Japanese question set as used in XIMERA [5].

2.3. HTS approach [7]

HTS consists of training and synthesis. An HMM voice built by the training part includes a set of statistical HMM models that represent context-dependent speech units. These units are organized as a set of decision trees for the state duration, fundamental frequency (F0), and cepstrum, as shown in Fig. 1. The leaf nodes of the decision trees include the model parameters that actually represent the mel-cepstrum, excitation, and duration of these units. At the speech synthesis phase, the input to the system is a full-context label sequence of the text to be synthesized. At first, the corresponding contexture-dependent HMMs are concatenated, then the state durations for the HMM sequence are determined in terms of the decision trees. After the speech parameters, including the mel-cepstral coefficients and F0 values at log-scale, are generated from the HMM sequence the output speech is synthesized by Mel Log Spectrum Approximation (MLSA) filtering [8].

3. Experiment method

3.1. Speech samples

Three datasets of speech samples are used in this paper. The first is web radio monologue speech from eight speakers (two females and six males) downloaded from the internet. The amount of speech samples for individual speakers varies from several minutes to hours. The second is 450 phone-balanced sentences uttered by two professional speakers (a female and a male) in a soundproof room, with the 900 utterances adopted from an in-house speech corpus. After this step, they are called clean speech. There are 31.99 and 34.36 minutes of speech samples from the female and male, respectively. The amount excludes the silence segments at the beginning and end of each utterance. We basically use the 900 utterances for the analysis experiments because their prompt text and phones were checked by humans. Hereafter, these are called correct transcripts for references. The third is several hours of read speech recorded in a soundproof room by the male speaker (clean speech). This subset enables us to make a common set for isolating undesirable influences.

3.2. Experiment setup

We conduct two experiments to investigate the influences of speech transcript errors on resultant HMM voices. Experiment 1 investigates the structures of decision trees for the state duration, F0, and cepstrum, focusing particularly on the numbers of leaf nodes and the mean tree depth from the root to the leaf nodes. Though the question sets used in the decision trees are important for this topic, we did not find a simple way to present them within the limited space. Experiment 2 examines the impact of speech transcript errors on HMM voice quality via a listening test. Basically, we simulate speech transcript errors through altering the beam search width of the LVASR when transcribing the speech into surface text and phone transcripts.

3.2.1. Experiment 1: Investigation of decision trees

Four sub-experiments are conducted for this purpose.
Experiment 1-A: Investigate relations between the numbers of leaf nodes of the decision trees and the amount of training data

- *Speech samples*: First and second sample sets
- *Beam search width*: Default setup in the LVASR
- *Building HMM voices*: Building HMM voices for individual speakers using their speech samples. For the

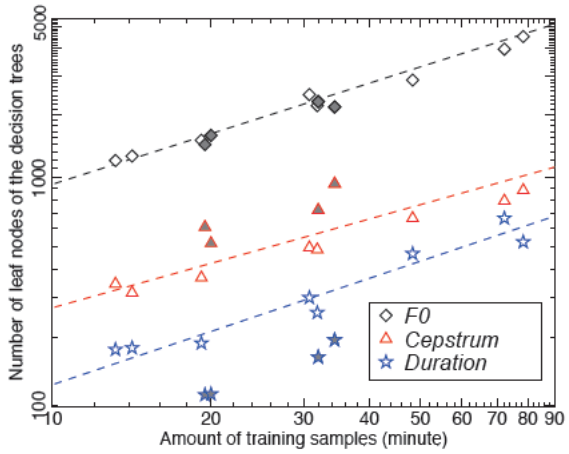

Figure 2: *Relations between the numbers of leaf nodes of the decision trees for state duration (indicated by stars), F0 (diamonds), as well as the cepstrum (triangles) and the amount of training data at log-scale. The empty symbols indicate the results for the web-radio monologue speech, and the filled symbols are the reference voices (built by the clean speech with correct transcripts). The dashed lines indicate the regression lines obtained from the results for the web-radio speech.*

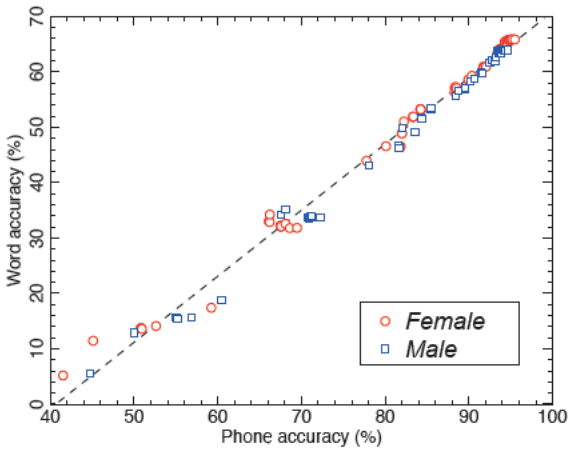

Figure 3: *Relations between word and phone accuracy occurring in the testbed. The dashed line indicates the regression line.*

clean speech, the correct transcripts are used for building two voices for each speaker: one with all clean speech (around 32 minutes) and the other with around a 19.5-minute randomly selected subset. After this, the four voices are called *reference voices*.

Experiment 1-B: Investigate relations between word accuracy and phone accuracy in this testbed

- *Speech samples*: Second sample set
- *Beam search width*: Varying from 60 to 130 by step 5

Experiment 1-C: Investigate influences of the speech transcript errors on the decision tree structures

- *Speech samples*: Second sample set
- *Beam width*: Typical beam width selected from *1-B*
- *Building HMM voices*: For each selected beam search width, we build two voices for each speaker: one with all 450 utterances (hereinafter *full sets*), and the other with a randomly selected subset of 19.5 minutes of data (*selected sets*), excluding the silent segments at the beginning and end of each utterance according to the phone-time alignment information. Note that some beam width values might be unreasonable as parts of the utterances are not fully recognized. In such cases, the non-recognized segments are included in the part of the silence segments at the beginning and end.

For reference, in the original dataset, there are 26,226 phones involved in the female's data and 26,467 phones in the male's data. With the changes of the beam search width used in the process of automatic speech transcription, the phone number of each resultant set differs. For the full sets, the mean phone number is 21,738.5 with standard deviation (SD) of 3,728 involved in the female's data, 22,089.3 with an SD of 2,905.99 in the male's data. For the selected sets, the mean phone number is 15,304.6 with an SD of 315.95 in the female's data, and 14,153.9 with an SD of 286.39 in the male's data. The statistical results show there may have some undesirable influences besides the transcript errors. For a comparison experiment, we make a common set from the third sample set to isolate the undesirable affecting factors.

Experiment 1-D: Comparison experiment to *Experiment 1-C*

- *Speech samples*: Third sample set
- *Beam width*: Typical beam width selected from *1-B*
- *Building HMM voices*: After transcribing the full third sample set, we build HMM voices from two selected subsets: a 22-min and 33-min set. The selected sets satisfy the following conditions.
  (1) All utterances are fully recognized and
  (2) they have phone accuracy around each of 65% to 95% by step 5%.
  Also, the 33-min set includes the 22-min set.

3.2.2. *Experiment 2: Listening test*

We perform a listening test to reveal the influences of transcript errors on HMM voice quality.

- *HMM voices*: 24 voices (for each speaker, 11 voices upon the selected sets obtained in *Experiment 1-C* plus a reference voice built by the 19.5-minute subset in *1-A*)
- *Stimuli*: 48 utterances (= 24 × 2 sentences; open test)
- *Listeners*: 4 (3 Japanese natives)
- *Score*: MOS on a five-point scale:
  1 (*bad*),
  2 (*poor*),
  3 (*fair*),
  4 (*good*), and
  5 (*excellent* or indistinguishable from human speech)
- *Condition*: The stimuli are played through headphones to the listeners in random order in a silent office.

4. Experiment results and observations

Figure 2 shows the experimental results on the relations between the amount of training data at log-scale and the numbers of leaf nodes of the decision trees for the state duration, F0, and cepstrum, respectively. The size of the decision trees measured as the number of their leaf nodes is "linearly" dependent on the amount of the training data used for building the web-based HMM voices (with certain speech transcript errors due to automatic speech transcription). In comparison with the reference voices (with the correct transcripts), the size of

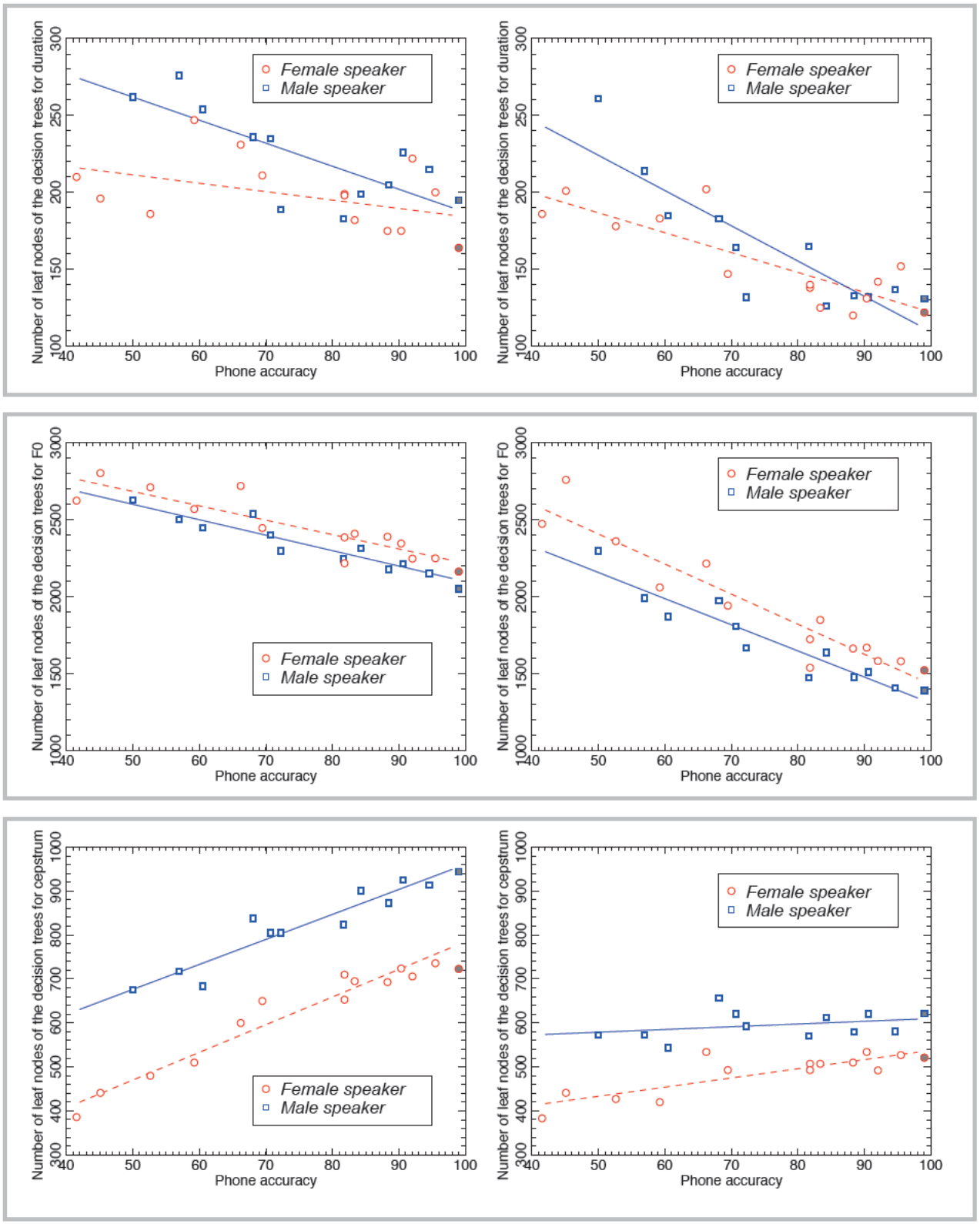

Figure 4: *Experimental results, achieved by using both the selected and full sets, on the impact of phone accuracy on the numbers of leaf nodes of the decision trees for the state duration (top box), F0 (middle box), and cepstrum (bottom box), respectively. The left panel in each box displays the results for the full sets, and the right for the selected sets. The filled symbols (squares and circles) indicate the results corresponding to the reference voices. The dashed and solid lines indicate the corresponding regression lines.*

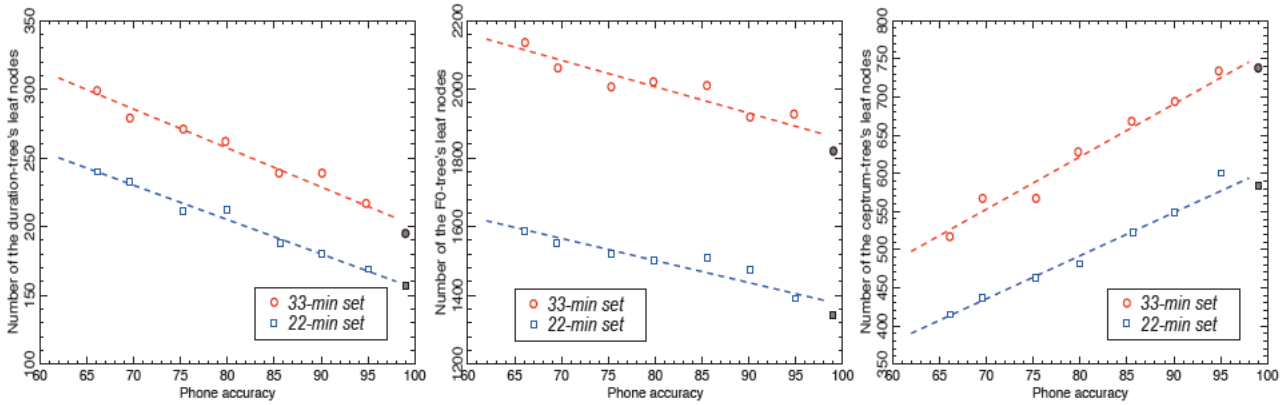

Figure 5: *Experimental results, achieved by using both 22-min and 33-min sets, on the impact of phone accuracy on the numbers of leaf nodes of the decision trees for the state duration (left panel), F0 (middle panel), and cepstrum (right panel), respectively.*

the decision trees for the web-based HMM voices is relatively large for the state duration and F0 but small for the cepstrum. This observation implies that some factors exist, such as speech transcript errors or noise, that affect the growth of the decision trees, compared to the natural growth appearing in the reference voices, given the amount of training samples.

Figure 3 displays the relations between word and phone accuracy occurring in this testbed. There is a linear relationship between the word and phone accuracy. After this, we only discuss the experiment results with the phone accuracy.

Figures 4 and 5 show the influences of the speech transcript errors on the numbers of leaf nodes of the decision trees of the state duration, F0, and cepstrum, achieved by *Experiments C and D*, respectively. The results indicate that high transcript errors (i.e., low phone accuracy) increase the number of leaf nodes of the decision trees of the state duration and F0, but decrease that of the cepstrum. When the phone accuracy nears 100%, the number of leaf nodes of all the decision trees approaches the natural one in the reference voices indicated by the filled symbols in the two figures. According to the results shown in Fig. 5, an increase in the amount of training sample sets consistently increases the number of leaf nodes of all the decision trees (state duration, F0, and cepstrum). Considering the non-consistence in Fig. 4, there are two undesirable factors affecting the experimental results apart from the speakers. One of these is the unrecognized segments that were labeled as part of the silences, and the other is the different contents involved in each set that were randomly selected. Comparing the results for the full set (right column) in Fig. 4 with the results in Fig. 5, both having the same utterances in each set, we can see that the unrecognized segments considerably affect the structures of the decision trees. It remains to be seen whether this influence is positive or negative for HMM voice quality.

Figure 6 displays the experimental results related to the mean depth of the decision trees for the HMM voices built by the 22-min and 30-min datasets. There seems to be no strong relation between the speech transcript errors and the mean depth of the decision trees. However, it is observable that the standard deviation of the tree depth is large when the phone accuracy is low, especially in the case of the F0 and cepstrum. This implies that the depth of individual branches of the decision trees varies considerably from one to another. That is, the structures of the decision trees are quite different. Also, according to the experimental results in Figs. 5 and 6, there is no clear relation between the number of leaf nodes of a tree and its mean depth; for example, the HMM voices shown in the filled symbols have different numbers of leaf nodes but quite similar mean depth and standard deviation.

Figure 7 shows the relations between the MOS and the phone accuracy of speech transcription of the speech data used for building the HMM voices. The filled circle and square in the upper-right corner indicate the results for the HMM voices with the correct transcripts. There is a clear relationship between the accuracies and corresponding HMM voice quality. The common utterances for each voice are 13.2% of the 19.5-minute training data. This may be the main reason for the MOS values locally waving in Fig. 7. From this experiment we can see that, in order to achieve *fair* voice quality at MOS value 3, phone accuracy must be over 80% and word accuracy above 50%.

5. Discussion

The experimental results clearly indicate that speech transcript errors have a considerable impact on HMM voices. Basically, the lower the speech transcript accuracy, the lower the HMM voice quality in the MOS. Speech transcript errors may show influences on HMM voices in the following three aspects: (1) phone deletion errors making the adjacent phone's duration longer; (2) phone substitution and insertion errors making a mismatch between phone labels and corresponding acoustic parameters; (3) word errors causing erroneous linguistic features involved in the full-context labels. All three aspects directly affect the accuracy of the linguistic features that are extracted from the surface text and phone transcripts. The linguistic features in turn affect the decision trees during the top-down clustering process directed by the linguistic feature-based question set. Also, our informal listening tests show that the speech rate of synthetic speech by the HMM models with lower phone accuracy sounds slow. However, phone accuracy seems to not show as significant influences on the continuity of synthetic speech as anticipated.

According to the experimental results displayed in Figs. 2, 4, and 5, the speech transcript errors increase the number of leaf nodes of the decision trees for both the state duration and F0, but suppress that of the decision trees for the cepstrum, when given the amount of the training data. It remains to be seen the mechanism by which speech transcript errors affect the growth of decision trees (measured as the number of leaf nodes) in a significantly different manner. Further work is needed on this aspect.

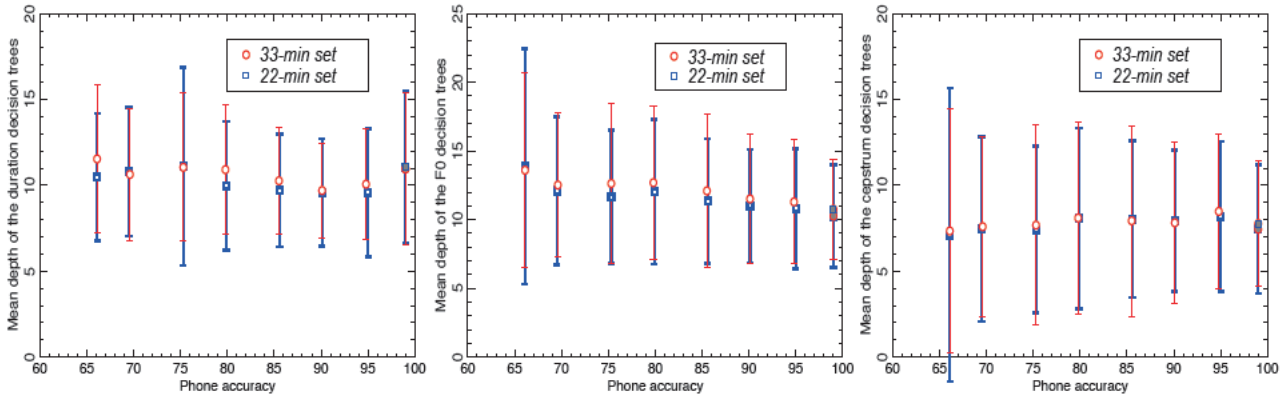

Figure 6: *Experimental results on the mean depth of the decision trees for the state duration (left panel), F0 (middle), and cepstrum (right) measured from the full sets when their phone accuracy is altered. The vertical bars indicate the standard deviations.*

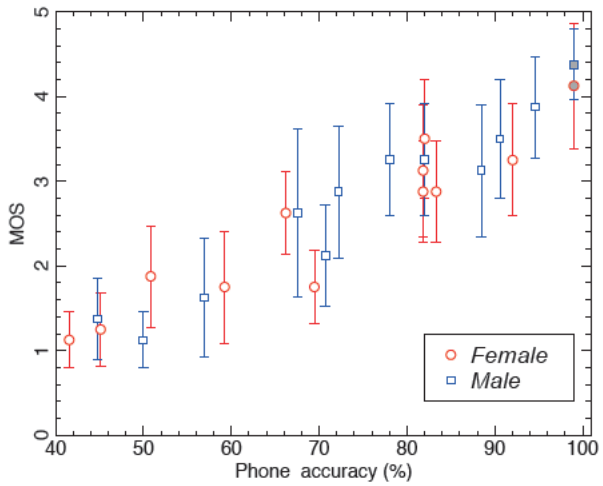

Figure 7: *Relations between HMM voice quality measured in mean opinion score (MOS) and phone accuracy of automatic transcription of the training samples (19.5 minutes per voice).*

6. Conclusion

This paper presented an investigation of the impact of speech transcript errors on HMM voices. The experimental results indicated that good speech transcription accuracy is very important for achieving high-quality HMM-based voices. In order to achieve fair voice quality, the phone and word accuracies of automatic speech transcription must be better than 80% and 50%, respectively. Our experiment results also indicated that speech transcript errors directly affected the growth of the decision trees. They increased the number of leaf nodes of the decision trees for the state duration and F0, but suppressed that of the decision trees for the cepstrum in comparison with the corresponding reference voices with correct transcripts. However, it remains to be seen why the speech transcript errors affect the growth of the decision trees in a significantly different manner. Further work is needed on this aspect.

7. Acknowledgements

This work was partly supported by a SCOPE fund. We are grateful for the valuable discussions and comments from the members of the speech synthesis team at NICT and the contributors of the audio contents on the internet.

8. References

[1] Nakamura S., *et al.*, "The ATR multi-lingual speech-to-speech translation system," IEEE Trans. on Speech and Audio Processing, 14 (2), 365–376, 2006.

[2] Tokuda K., Kobayashi T., and Imai S., "Speech parameter generation from HMM using dynamic features," *Proc. IEEE ICASSP*, vol. 1, pp. 660–663, 1995.

[3] http://hts.ics.nitech.ac.jp/?HTS

[4] Ni J. and Kawai H., "An unsupervised approach to creating web audio contents-based HMM voices," *to appear at Interspeech2010*.

[5] Kawai H. *et al.*, "XIMERA: a new TTS from ATR based on corpus-based technologies," *5th ISCA Speech Synthesis Workshop*, pp. 179–184, 2004.

[6] Toda T. and Tokuda K., "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *Proc. Interspeech (Eurospeech)*, pp. 2801-2804, 2005.

[7] Tokuda K., Zen H., Black A.W., "An HMM-based approach to multilingual speech synthesis," *Text to Speech Synthesis: New Paradigms and Advances*, Prentice Hall, pp. 135–153, 2004.

[8] Fukada T., Tokuda K., Kobayashi K., and Imai S., "An adaptive algorithm for mel-cepstral an analysis of speech," *Proc. IEEE ICASSP*, vol. 1, pp. 137–140, 1992.