Thousands of voices for HMM-based speech synthesis

J. Yamagishi, Bela Usabaev, Simon King, Oliver Watts, John Dines, Jilei Tian, Rile Hu, Yong Guan, Keiichiro Oura, Keiichi Tokuda, Reima Karhila, and Mikko Kurimo. Thousands of voices for HMM-based speech synthesis. In Proc. Interspeech, pages 420-423, Brighton, U.K., September 2009.

Thousands of Voices for HMM-based Speech Synthesis

Junichi Yamagishi[1], *Bela Usabaev*[2], *Simon King*[1], *Oliver Watts*[1], *John Dines*[3], *Jilei Tian*[4], *Rile Hu*[4], *Keiichiro Oura*[5], *Keiichi Tokuda*[5], *Reima Karhila*[6], *Mikko Kurimo*[6]

[1]University of Edinburgh [2]Universität Tübingen [3]Idiap Research Institute
[4]Nokia [5]Nagoya Institute of Technology [6]Helsinki University of Technology

jyamagis@inf.ed.ac.uk

Abstract

Our recent experiments with HMM-based speech synthesis systems have demonstrated that speaker-adaptive HMM-based speech synthesis (which uses an 'average voice model' plus model adaptation) is robust to non-ideal speech data that are recorded under various conditions and with varying microphones, that are not perfectly clean, and/or that lack of phonetic balance. This enables us consider building high-quality voices on 'non-TTS' corpora such as ASR corpora. Since ASR corpora generally include a large number of speakers, this leads to the possibility of producing an enormous number of voices automatically. In this paper we show thousands of voices for HMM-based speech synthesis that we have made from several popular ASR corpora such as the Wall Street Journal databases (WSJ0/WSJ1/WSJCAM0), Resource Management, Globalphone and Speecon. We report some perceptual evaluation results and outline the outstanding issues.

Index Terms: speech synthesis, HMMs, speaker adaptation

1. Introduction

Statistical parametric speech synthesis based on hidden Markov models (HMMs) [1] is now well-established and can generate natural-sounding synthetic speech. In this framework, we have pioneered the development of the HMM Speech Synthesis System, HTS (H Triple S) [2]. In conventional speech synthesis including HTS, large amounts of phonetically-balanced speech data recorded in highly-controlled recording studio environments are typically required to build a voice. Although using such data is a straightforward solution for high quality synthesis, the number of voices available will always be limited, because recording costs are high.

On the other hand, our recent experiments with HMM-based speech synthesis systems have demonstrated that speaker-adaptive HMM-based speech synthesis (which uses an 'average voice model' plus model adaptation) is robust to non-ideal speech data that are recorded under various conditions and with varying microphones, that are not perfectly clean, and/or that lack of phonetic balance[Add references]. This enables us consider building high-quality voices on 'non-TTS' corpora such as ASR corpora. Since ASR corpora generally include a large number of speakers, this leads to the possibility of producing an enormous number of voices automatically.

In this paper we explain the thousands of voices for HMM-based speech synthesis that we have made from several popular ASR corpora such as the Wall Street Journal databases (WSJ0/WSJ1/WSJCAM0), Resource Management, Globalphone and Finnish and Mandarin Speecon. We will report some analysis results, perceptual evaluation results, an application, and outline the outstanding issues of the voices.

2. HTS voices trained on ASR corpora

2.1. Framework of TTS systems

All TTS systems are built using the framework from the "HTS-2007 / 2008" system ([3]), which was a speaker-adaptive system entered for the Blizzard Challenge 2007 and 2008 ([4]).

2.2. ASR speech databases used for TTS systems

In conventional speech synthesis research, phonetically-balanced speech databases are typically used. A phonetically-balanced dataset (e.g., complete diphone coverage) is required for each individual speaker, since conventional systems are speaker-dependent. In multi-speaker sets of speech synthesis data (e.g., CMU-ARCTIC[1]), it is common for the same set of phonetically-balanced sentences to be re-used for each speaker. Therefore, pooling the data from multiple speakers does not always significantly increase phonetic coverage. Compared to this, the sentences chosen for ASR corpora tend to be designed to achieve phonetic balance across multiple speakers, or simply chosen randomly. Therefore, phonetic coverage increases with the number of speakers. However, each individual speaker typically records a very limited number of utterances (e.g., fewer than 100). Building TTS voices from these ASR corpora is in itself a new challenge.

We hypothesised that it would be feasible to build speaker-adaptive HTS systems using ASR corpora, since adaptive training techniques (e.g., SAT) can normalize speaker differences, and since the total phonetic coverage of ASR corpora may be better than that of TTS (see Section 2.4). Therefore we used a number of popular ASR corpora such as the Wall Street Journal databases (WSJ0/WSJ1/WSJCAM0), Resource Management, Globalphone, Finnish and Mandarin Speecon, and Japanese JNAS.

The Wall Street Journal corpus (WSJ) is particularly well-suited to this since it provides a large quantity of transcribed read speech data of mostly good quality (though not in the same category as purpose-built speech synthesis databases). Thus the WSJ0 was the primary corpus used for comparison of speaker-dependent and speaker-adaptive HMM-based TTS systems. The speaker-dependent systems were built from the subset called "very long term" which includes about 2,400 sentences per speaker for a small number of speakers. Average voice models were built using other subsets: short term, long term (excluding the speakers from very long term), develop-

[1]A free database for speech synthesis, http://festvox.org/cmu_arctic/

Table 1: Triphone coverage of ASR and TTS corpora

Name	triphones/speaker	triphones/corpus
CMU-ARCTIC	10041	10708
WSJ0 (short/SI-84)	3287	18577
WSJ0+1 (short/SI-284)	4220	23776
WSJCAM0 (total)	3036	23534
RM (ind_total)	1091	7162

ment, and evaluation. In total, 110 speakers utter from 80 to 600 sentences each. We compared speaker-dependent models trained with a reasonably large amount of data (2,400 sentences – which is twice the size of a single-speaker CMU-ARCTIC dataset) against various speaker-adaptive systems.

The speecon corpora includes speech data recorded in various amounts of background noise (e.g., "car" or "public spaces"). Although it may eventually be possible to use such data for speech synthesis, we chose a set of speech data recorded in relatively quiet "office" environments (although this is not still perfectly clean). The data includes isolated word or spelling pronunciation utterances and phonetically balanced sentences. Since we are unsure of the effects of using large quantities of isolated word or spelling pronunciation utterances on synthesis, we used only the phonetically balanced sentences as training sentences for the average voice model in this experiment.

2.3. Front-end processing

The labels for the data were automatically generated from the word transcriptions and speech data using the Unisyn lexicon [5] and Festival's Multisyn Build modules for English and Spanish voices, and using Nokia's in-house lexica and TTS modules for Finnish and Mandarin voices, with no further modification. The multisyn Build modules identified utterance-medial pauses, vowel reductions, or reduced vowel forms and they were added to the labels. For the out-of-vocabulary words, letter-to-sound rules of the Festival's Multisyn were used. English and Spanish phonesets are based on IPA and Finnish and Mandarin phonesets are based on SAMPA-C.

2.4. Analysis of ASR corpora from TTS point of view – phonetic coverage

Triphone coverage is a simple way to measure the phonetic coverage of a corpus. Table 1 shows the average number of different triphone types per speaker and the total number of different triphone types in the various corpora. A larger number of types implies that the phonetic coverage is better, which in turn implies that the corpus is more suitable for speech synthesis. For comparison, the triphone coverage of the CMU-ARCTIC speech database which includes four male and two female speakers is also shown.

We can see although the average number of triphone types for each speaker in the CMU-ARCTIC database is clearly larger than for any single speaker from an ASR corpus, the total triphone coverage across all speakers in the CMU-ARCTIC database is about the same (because all speakers say the same set of sentences). In contrast, the triphone coverage of the complete WSJ0, WSJ1 and WSJCAM corpora is much higher than CMU-ARCTIC. This leads us to believe that these ASR corpora should be better for building speaker-independent/adaptive HMM-based TTS systems as well as speaker-independent ASR systems. The RM corpus, because of its very limited domain and small word vocabulary, has relatively poor coverage and would be unsuitable for use as a TTS corpus unless combined with other data.

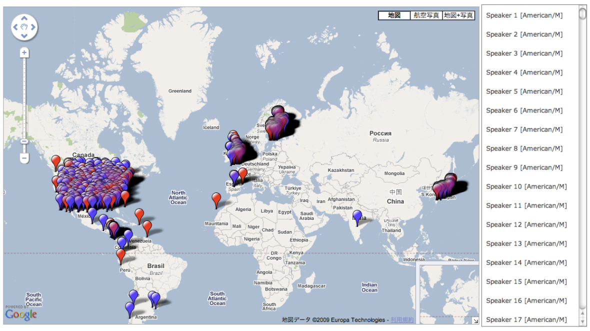

Figure 1: Geographical representation of HTS voices trained on ASR corpora for EMIME projects. Blue markers show male speakers and red markers show female speakers. Available online via http://www.emime.org/learn/speech-synthesis/listen/Examples-for-D2.1

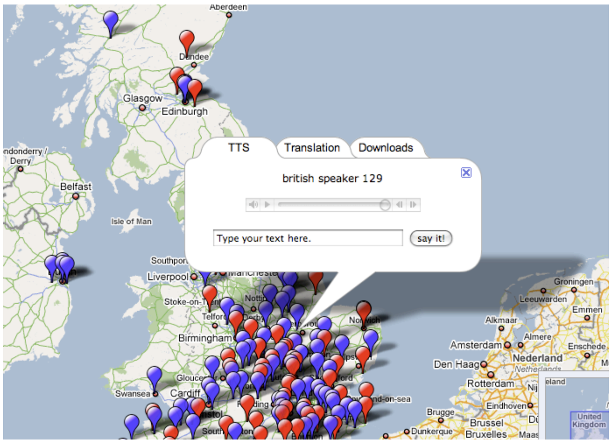

Figure 2: All English HTS voices can be used as online TTS on the geographical map.

2.5. Demonstration of the HTS voices

We built speaker-adaptive HMM-based TTS systems from each corpora above and adapt them to all speakers available. Informal listening revealed that there are a few speakers whose synthetic speech sounds worse than other speakers. This may be because the available for these speakers has poor phonetic coverage or because of other factors such as properties of the speaker's voice or the recording quality. This will be investigated in future work. The phenomenon is analogous to the familiar situation in ASR, where WER varies widely across some speakers and is especially high for a small number of speakers. Samples are available from http://www.emime.org/learn/speech-synthesis/listen/Examples-for-D2.1

2.6. Geographical representation and online demo

One of important advantages of using ASR corpora is the large number of speakers. Building TTS voices on such data allows the creation of many more voices than has previously been possible for TTS. In fact, we built so many voices (1500+ including some voices built outside the EMIME project but using the same techniques, which we believe is the largest known collection of synthetic voices in existence) it became impossible to represent them in list or table form. Instead, we devised an interactive geographical representation, shown in Figure 1. Each marker corresponds to an individual speaker. Blue markers show male speakers and red markers show female speakers. Some markers are in arbitrary locations (in the correct country) because precise location information is not available for all speakers. This geographical representation, which includes an interactive TTS demonstration of many of the voices, is available from the URL provided. Clicking on a marker will play synthetic speech from that speaker[2]. As well as being a convenient interface to compare the many voices, the interactive map is an attractive and easy-to-understand demonstration of the technology being developed in EMIME.

2.7. Multidimensional scaling of male speakers included in WSJ0 corpus

Another way to visualize the speakers is to place them not in a geographical space, but in a space derived from properties of the speech. This can be achieved using multidimensional scaling [6]. We generated a set of speech samples from all the HTS voices trained on the WSJ0 corpus using all test sentences from the Blizzard Challenge 2008. We then calculated the average mel-cepstral distance between the speech for all pairs of voices, placing the values in a mel-cepstral distance table. For simplicity, the unadapted duration models of the average voice model were used so that the number of frames of synthetic speech for each speaker is same. Then we applied a classic multidimensional scaling technique [6] to the mel-cepstral distance table and examined the resulting two-dimensional space, which is shown in Figure 3.

The axes of this space do not have any meaning, but MDS attempts to preserve the pairwise distances between speakers given in the mel-cepstral distance table. In other words, similar speakers will be close to one another in this space. For example, speakers 012, 01e, 029, 02b and 021 are similar to one other (in terms of mel-cepstral distance) and speakers 22h, 422, and 423 are relatively different from other speakers. We can only use very few target speakers in formal listening test, so it is important to investigate the distribution of speakers in other ways, such as MDS.

3. Evaluation

In this experiment, we confirmed that our speaker-adaptive systems built on ASR corpora show the same tendencies as those previous systems. We also confirmed that our speaker-adaptive systems provide good baseline performance.

3.1. Average voice model training data

We built two kinds of average voice model. The first was built using 50 utterances per training speaker ("condition 1"). If a speaker has more than 50 utterances, a subset of 50 was chosen randomly. The second average voice model was built using all available utterances from all training speakers ("condition 2"). The numbers of training sentences are 2950 and 10847 sentences for male average voice models in conditions 1 and 2 respectively, and for female average voice models there are

[2]Currently the interactive mode supports English and Spanish only. For other languages this only provides pre-synthesised examples, but we plan to add an interactive type-in text-to-speech feature in the near future.

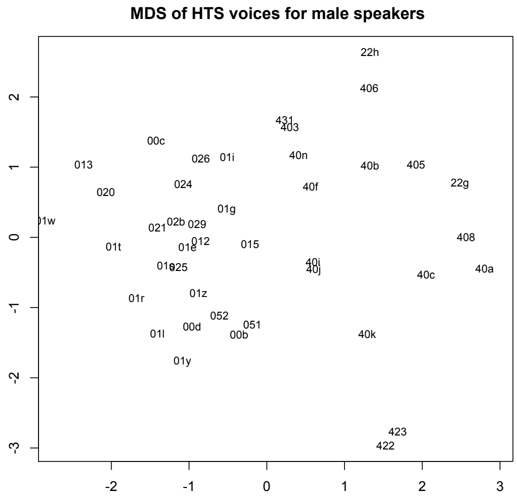

Figure 3: Multidimensional scaling of HTS voices trained on WSJ0 corpus. The three characters at each point correspond to the name of each speaker in the database.

3000 and 12151 sentences respectively. They have 5.7 hours, 21.1 hours, 5.9 hours, and 24.6 hours of speech duration, respectively. By providing a part of training data for speaker-dependent models to the average voice models, we compared the speaker-adaptive systems with speaker-dependent systems.

3.2. Speaker-dependent model training data

To examine the effect of corpus size, three speaker-dependent systems were built, using 100 randomly chosen sentences (about 6 minutes in duration), 1000 randomly chosen sentences (about 1 hour in duration) and 2000 randomly chosen sentences (about 2 hours in duration) respectively from the target speaker.

3.3. Objective evaluation

Table 2 shows the objective measures for each system. From the results for speaker 001, we can confirm that the speaker-adaptive systems using all available average voice model training data ("condition 2") outperform the speaker-adaptive systems using an equal amount of speech data per training speaker ("condition 1"). In addition, as we expected, we can see that when the amount of target speaker speech data is less than about 1 hour, speaker-adaptive systems outperform speaker-dependent systems. Once the amount of speech data is more than about 1 hour, speaker-dependent systems start to become better than speaker-adaptive systems. This result is consistent with previous results.

The RMSE of $\log F_0$ for the speaker 002 shows unexpected tendencies. All the systems using 2 hours of target speaker speech data have worse RMSE than those using 1 hour of data. A possible explanation for this is that the speaker's speaking style was not consistent over the long-term recording sessions (e.g., the average value and range of F_0 varied session by session). This may be investigated in future work: although the EMIME application may operate with less target speaker data than this, there may still be multiple speech capture sessions as the device is used on different occasions. We chose the male speaker 001 as the target speaker for the subjective (listening test) evaluation.

Table 2: The objective measures of each speaker-dependent (SD) and speaker-adaptive (SA) systems built using various amounts of speech data from the target speaker. Underlined figures indicate the best performing system under each objective measure for each target speaker (i.e., in each column). MCD and $\log F_0$ show mel-cepstral distance and RMSE of $\log F_0$, respectively.

(a) 6 minutes of target speaker data

	Speaker 001		*Speaker 002*	
	MCD	$\log F_0$	MCD	$\log F_0$
System	(dB)	(cent)	(dB)	(cent)
SD	9.05	407	7.18	195
SA (condition 1)	5.46	393	4.97	168
SA (condition 2)	5.38	369	5.09	186

(b) 1 hour of target speaker data

	Speaker 001		*Speaker 002*	
	MCD	$\log F_0$	MCD	$\log F_0$
System	(dB)	(cent)	(dB)	(cent)
SD	5.27	354	4.86	174
SA (condition 1)	5.36	398	4.99	176
SA (condition 2)	5.25	352	4.98	174

(c) 2 hours of target speaker data

	Speaker 001		*Speaker 002*	
	MCD	$\log F_0$	MCD	$\log F_0$
System	(dB)	(cent)	(dB)	(cent)
SD	5.18	348	4.83	190
SA (condition 1)	5.32	386	4.97	180
SA (condition 2)	5.25	351	4.97	182

3.4. Subjective evaluation

We adopted the evaluation methods used in the Blizzard Challenge 2008. English synthetic speech was generated for a set of 600 test sentences, including 400 sentences from conversational, news and novel genres (used to evaluate naturalness and similarity) and 200 semantically unpredictable sentences (used to evaluate intelligibility). A subset of these sentences were then chosen randomly for use in the listening test (the exact number required depends on the number of systems being compared — see [4] for details of the Latin Square experimental design.) The number of listeners for this experiment was 26.

Figure 4 shows the results. The perceptual evaluation reveal the same tendencies as the objective evaluations. The speaker-adaptive systems using the all the data ("condition 2") were found by listeners to be better in terms of naturalness and similarity than the speaker-adaptive systems using an equal amount of speech data. We can again see that when the amount of speech data is less than about 1 hour, speaker-adaptive systems outperform speaker-dependent systems in every way. Once the amount of speech data is about 1 hour, the speaker-dependent system and speaker-adaptive system in condition 2 have almost the same scores. When the amount of speech data is about 2 hours, the speaker-dependent system starts to have better naturalness and intelligibility than the speaker-adaptive system.

These results are consistent with previous analyses. We conclude that the performance of our baseline speaker-adaptive system is good and comparable to other state-of-the-art HMM-based speech synthesis systems.

4. Conclusions

Building TTS voices on ASR speech database allows the creation of many more voices than has previously been possible for TTS. We have shown their analysis/evaluation results and applications using a geographical map. These voices would have potential for some applications such as medical voice banking or virtual game such as second life. Our future work is to analyze the difference of the quality of the voices.

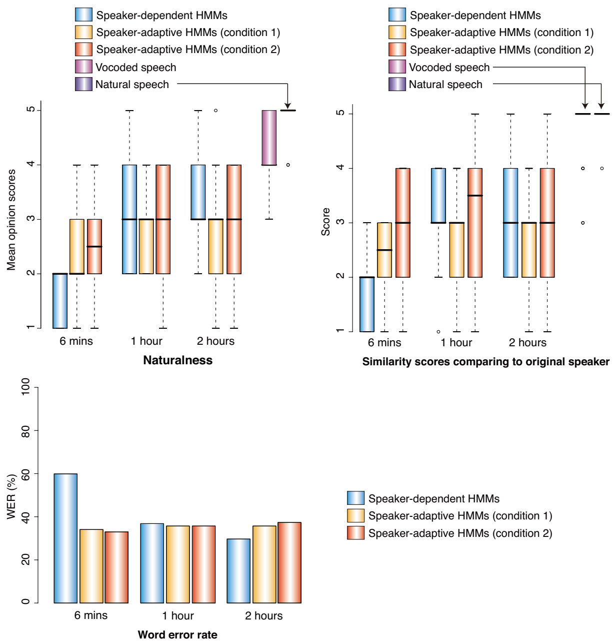

Figure 4: Subjective evaluation results for speaker-dependent and speaker-adaptive HMM-based TTS systems built on ASR corpora.

Acknowledgements The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project `http://www.emime.org`). JY is partially supported by EPSRC. SK holds an EPSRC Advanced Research Fellowship. BU was supported by ERASMUS Konsortium KOOR/BEST. This work has made use of the resources provided by the Edinburgh Compute and Data Facility which is partially supported by the eDIKT initiative (`http://www.edikt.org.uk`).

5. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH-99*, Sept. 1999, pp. 2374–2350.

[2] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A. Black, and T. Nose, *The HMM-based speech synthesis system (HTS) Version 2.1*, http://hts.sp.nitech.ac.jp/.

[3] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "A robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. Speech, Audio & Language Process.*, 2009, (in press).

[4] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *Proc. Blizzard Challenge Workshop*, Brisbane, Australia, September 2008.

[5] S. Fitt and S. Isard, "Synthesis of regional English using a keyword lexicon," in *Proc. Eurospeech 1999*, vol. 2, Sept. 1999, pp. 823–826.

[6] T. Cox and M. Cox, *Multidimensional Scaling*. Chapman and Hall, 2001.