Thousands of Voices for HMM-Based Speech Synthesis–Analysis and Application of TTS Systems Built on Various ASR Corpora

Junichi Yamagishi, *Member, IEEE*, Bela Usabaev, Simon King, *Senior Member, IEEE*, Oliver Watts, John Dines, *Member, IEEE*, Jilei Tian, Yong Guan, Rile Hu, Keiichiro Oura, Yi-Jian Wu, Keiichi Tokuda, *Member, IEEE*, Reima Karhila, and Mikko Kurimo, *Senior Member, IEEE*

***Abstract*—In conventional speech synthesis, 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 phonetic balance. This enables us to 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 demonstrate the thousands of voices for HMM-based speech synthesis that we have made from several popular ASR corpora such as the Wall Street Journal (WSJ0, WSJ1, and WSJCAM0), Resource Management, Globalphone, and SPEECON databases. We also present the results of associated analysis based on perceptual evaluation, and discuss remaining issues.**

***Index Terms*—Automatic speech recognition (ASR), average voice, hidden Markov model (HMM)-based speech synthesis, H Triple S (HTS), speaker adaptation, speech synthesis, SPEECON database, voice conversion, WSJ database.**

Manuscript received May 11, 2009; revised February 08, 2010. Current version published June 16, 2010. This work was supported by the European Community's Seventh Framework Program (FP7/2007-2013) under Grant Agreement 213845 (the EMIME project http://www.emime.org). This work has made use of the resources provided by the Edinburgh Compute and Data Facility which is supported in part by the eDIKT initiative (http://www.edikt.org.uk). The work of J. Yamagishi was supported in part by EPSRC. S. King holds an EPSRC Advanced Research Fellowship. The work of B. Usabaev was supported by ERASMUS Konsortium KOOR/BEST. Simplified descriptions of this research are introduced in the conference proceeding for Interspeech 2009 [1]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Olivier Rosec.

J. Yamagishi, S. King, and O. Watts are with the Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh EH8 9AB, U.K. (e-mail: jyamagis@inf.ed.ac.uk; simon.king@ed.ac.uk; o.s.watts@sms.ed.ac.uk).

B. Usabaev is with the Universität Tübingen, 72074 Tübingen, Germany (e-mail: belausabaev@googlemail.com).

J. Dines is with the Idiap Research Institute, CH-1920, Martigny, Switzerland (e-mail: john.dines@idiap.ch).

J. Tian, Y. Guan, and R. Hu are with the Nokia Research Center, Beijing, 100176, China (e-mail: jilei.tian@nokia.com; rile.hu@nokia.com; ext-yong.guan@nokia.com).

K. Oura and K. Tokuda are with the Department of Computer Science and Engineering, Nagoya Institute of Technology (NIT), Nagoya 466-8555, Japan (e-mail: uratec@sp.nitech.ac.jp, tokuda@nitech.ac.jp).

Y-J. Wu was with the Nagoya Institute of Technology (NIT), Nagoya, 466-8555, Japan. He is now with the TTS Group, Microsoft Business Division, Beijing 100084, China (e-mail: yijiwu@microsoft.co).

R. Karhila and M. Kurimo are with the Adaptive Informatics Research Centre, Helsinki University of Technology, FIN-02015 TKK, Finland (e-mail: rkarhila@james.hut.fi; Mikko.Kurimo@tkk.fi).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2045237

I. Introduction

STATISTICAL parametric speech synthesis based on hidden Markov models (HMMs) [2] 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, H Triple S (HTS) [3], [4].

In the conventional speech synthesis framework 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 the costs associated with recording and manually annotating speech data are high.

Another practical, but equally important, reason is footprint available for text-to-speech (TTS) synthesis systems. In general, disk space available for TTS systems in commercial products is limited, and thus it is infeasible for the systems to have a large variety of voices since the number of voices is a factor of footprint.

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 phonetic balance [5], [6]. In [6], a high-quality voice was built from "found" audio, freely available on the web. These data were not recorded in a studio and had a small amount of background noise. The recording condition of the data was not consistent: the environment and microphone also varied. This enables us to consider building high-quality voices on other "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 addition, speaker-adaptive HMM-based speech synthesis is efficient in the sense of footprint. Compared with so-called unit-selection synthesis, the footprint of HMM-based speech synthesis systems is usually smaller because we store statistics of acoustic subword models rather than templates of the subword units [7]. Furthermore, the statistics of the subword models in the speaker-adaptive HMM-based speech synthesis (i.e., average voice models) is speaker-independent and thus can be shared among an arbitrary group of speakers. The speaker-dependent footprint is only a set of transforms used for speaker adaptation, which is usually much smaller than the statistics of the subword models since the transforms are further shared among the subword models.

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 [8], and a Cambridge version of WSJ0 called WSJCAM0 [9]), Resource Management (RM) [10], GlobalPhone [11] and Finnish and Mandarin SPEECON [12]. We believe these voices form solid benchmarks and provide good connections to ASR fields. This paper also reports a series of analysis results for investigating the effect of such non-ideal data from a TTS perspective, suggests useful applications for the thousands of voices, and addresses outstanding issues.[1]

This paper is organized as follows. Section II gives an overview and analysis of the ASR corpora used for building TTS systems. A brief overview of speaker-adaptive HMM-based speech synthesis, details of voices built, and applications which make use of thousands of TTS voices are also given. Section III introduces evaluation methodologies used. Sections IV and V describe analysis of the use of ASR corpora for TTS. Then Section VI concludes the paper by briefly summarizing our findings.

II. TTS Voices Trained on ASR Corpora

A. TTS and ASR Speech Databases

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[2]), it is common for the same set of phonetically balanced sentences to be reused 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 are 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 utterances).

However, we hypothesized that it would be feasible to build speaker-adaptive HTS systems using ASR corpora, since adaptive training techniques (e.g., SAT [13]) can normalize speaker differences, and since the total phonetic coverage of ASR corpora may be better than that of TTS (see Section II-H). Therefore, we used a number of recognized, publicly available ASR corpora—the Wall Street Journal databases (WSJ0, WSJ1, and WSJCAM0), Resource Management (RM), GlobalPhone, Finnish and Mandarin SPEECON, and Japanese JNAS [14]. The subsections which follow overview the ASR databases in each language.

TABLE I
Details of English ASR Corpora Used for Building HMM-Based TTS Systems

Corpus (subset)	Speakers	Sentences/speaker	Sentences
RM (ind_train)	80	40	3200
RM (ind_dev)	40	40	1600
RM (ind_eavl)	40	40	1600
RM (ind_total)	160	40	6400
RM (dep_dev)	12	100	1200
RM (dep_eval)	12 (dep_dev)	100	1200
RM (dep_total)	12	200	2400
WSJ0 (short)	84	86.1	7236
WSJ0 (long)	12	600	7201
WSJ0 (very long)	3 (long)	2400	7199
WSJ0 (dev)	10	194.8	1948
WSJ0 (eval)	8	163.4	1307
WSJCAM0 (train)	92	85.4	7861
WSJCAM0 (dev)	20	73.5	1471
WSJCAM0 (eval)	28	74.1	2076
WSJCAM0 (total)	140	81.5	11408
WSJ1 (short)	200	191.3	38278
WSJ1 (long)	25	1241.6	31029

B. English ASR Speech Databases

Table I gives detailed information on the number of speakers and sentences included in predefined subsets of the English ASR corpora. No speakers (except a very limited number of speakers included in the subsets called "very long" in WSJ0 or "long" in WSJ1) have a sufficient number of sentences to train HMMs that can be used for TTS systems. For the training of speaker-dependent HMMs, we usually require over five hundred sentences. Therefore building TTS voices from these ASR corpora is, in itself, a new challenge.

Since the English corpora provide varying quantities of transcribed read speech data of mostly good quality (though not in the same category as purpose-built speech synthesis databases), they were used for 1) comparison of speaker-dependent and speaker-adaptive HMM-based TTS systems, 2) analysis of the effect of the quantity of data used for the average voice models, and also 3) comparison of footprints of acoustic models built. These topics are mentioned in Sections IV, V-C, and II-K, respectively.

The WSJ0 was particularly well-suited for the 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 2400

[1]In this paper, we do not explore multilingual or cross-lingual approaches to acoustic modeling. We simply use language-dependent acoustic models and TTS systems.

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

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), development, 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 (2400 sentences—twice the size of a single-speaker CMU-ARCTIC dataset) with various speaker-adaptive systems.

For the analysis of the effect of the quantity of speech data used for training the average voice models, and comparison of footprints of the acoustic models built, we used speaker-independent subsets (short, train, or ind_train in Table I) of the RM, WSJ0, WSJCAM0, and WSJ1 databases and built average voice models on each database. The total amounts of speech data of the subsets for RM, WSJ0, WSJCAM0, and WSJ1 are 5 hours, 15 hours, 22 hours, and 66 hours, respectively in terms of duration including silences and pause.

C. Finnish and Mandarin SPEECON Databases

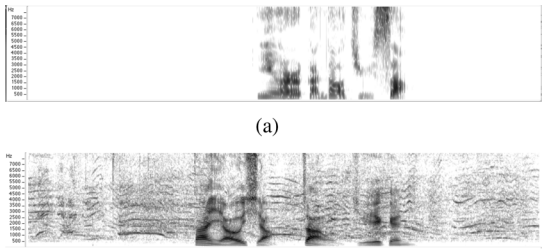
For the Finnish and Mandarin average voice models, we have used the SPEECON "Speech Databases for Consumer Devices" [12]. The SPEECON databases include speech data recorded in various conditions with various amounts of background noise, detailed below. We directly cite definitions of the noise categories from [12]:

Office	mostly quiet; if background noise is present, it is usually more or less stationary.
Entertainment	a home environment but noisier than office; the noise is more coloured and non-stationary; it may contain music and other voices.
Public place	may be indoor or outdoor; noise levels are hard to predict.
Car	a medium to high noise level is expected of both stationary (engine) and instantaneous nature (wipers).

Sample spectrograms for Mandarin speech recorded in office and public space environments are shown in Fig. 1. Noise levels in dB [A] for each environment included in Mandarin SPEECON are shown in Table II. We can see that the public space and car environments have larger means and variances. Details of each environment are shown in Table III. The lengths of speech data recorded in office, public space, entertainment, and car are 12.3 hours, 11.3 hours, 4.9 hours, and 5.2 hours, respectively. The structure of the Finnish SPEECON database is identical to that of the Mandarin one.

Since the SPEECON corpora provide speech data recorded in various conditions, they were used for 1) comparison with purpose-built perfectly clean high-quality speech synthesis databases, and 2) analysis of the effect of inconsistent recording conditions. These topics are mentioned in Section V-D together.

For the analysis of the effect of the inconsistent recording conditions, we chose a set of speech data recorded in the relatively quiet "office" environments (although this is not still perfectly clean: see Max value!) for training the baseline system



(a)

(b)

Fig. 1. Spectrograms of clean and noisy data included in the Mandarin SPEECON database. (a) Clean data recorded in office space. (b) Noisy data recorded in public space.

TABLE II
NOISE LEVEL IN DB [A] FOR EACH ENVIRONMENT INCLUDED IN THE MANDARIN SPEECON DATABASE

	Noise dB [A]			
Environment	Mean	Variance	Min	Max
Office	44.7	25.5	34	54
Public space	56.7	45.3	41	73
Entertainment	46.9	24.2	37	61
Car	57.0	130.0	34	71

TABLE III
DETAILS OF THE MANDARIN SPEECON CORPUS USED FOR BUILDING AVERAGE VOICE MODELS. THE FINNISH SPEECON CORPUS ALSO HAS THE SAME STRUCTURE

Environment	Speakers	Sentences/speaker	Sentences
Office	200	29.6	5916
Public space	180	29.9	5378
Entertainment	75	29.9	2240
Car	75	30.0	2247
Total	530	29.8	15781

and compared it with a system using all data regardless of the environment. Note that the system has about three times as much speech data as the baseline system. If the amount of noisy data is equal to that of clean speech data, then clearly the TTS voices adapted from the model trained on the noisy data will be worse than those from the model trained on clean data. We therefore analyze the advantages (and disadvantages) of the more likely situation, where much more noisy data is available than clean data.

For the comparison with purpose-built perfectly clean speech synthesis databases, we utilized the systems above and a normal TTS system trained on phonetically balanced speech data recorded in highly controlled recording studio environments.

The databases also include 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 phonetically balanced sentences as training sentences for the average voice model in this experiment.

D. Japanese JNAS and Spanish Globalphone Databases

The Japanese Newspaper Article Sentences database (JNAS) contains speech recordings and their orthographic transcriptions of 306 speakers (153 males and 153 females) reading excerpts

from the Mainichi Newspaper and the ATR 503 phonetically balanced sentences [14]. From the database, we randomly chose 50 female and 50 male speakers, who have a total of 14 134 utterances, as training speakers for the Japanese average voice models. The length of the chosen speech data is 19.5 hours.

The GlobalPhone database is a multilingual speech and text database that covers 15 languages [11]. From the database, we utilized all of the one hundred Spanish speakers, who have a total of 6620 utterances, as training speakers for the Spanish average voice models. The length of the chosen speech data is 22 hours.

The Japanese and Spanish models were utilized for demonstrating an application using many TTS voices mentioned in Section II-M.

E. Phonesets, Lexica, and Front-End Processing

Contrary to normal TTS databases where professional or semiprofessional narrators utilize standard accents and speaking styles, the speakers included in the ASR databases have a variety of accents. For instance, the Mandarin SPEECON database is made up equally of four major dialectal accents (Beijing, Chongqing, Shanghai, and Provinces).

Using speech recordings that comprised a variety of accents for training could prove either advantageous or disadvantageous. If the target speaker has an accent for which training data is not available, models trained on the various accents would be more appropriate since they have larger variance and can capture the variation in the unseen accent. On the other hand, when the target accent is limited to, for example the British received pronunciation (RP) accent, as it is in the Blizzard Challenge, a more appropriate average voice model would be one trained only on RP speakers, rather than one trained on various accents.

Since the Unilex pronunciation lexicon [16] from CSTR supports multiple accents of English in a unified way—by deriving surface-form pronunciations from an underlying meta-lexicon defined in terms of key symbols—it is possible, in theory, to prepare different phonesets for each accent. The same framework may be used for accents in other languages. In practice, however, time constraints meant we were unable to do this, and we simply used an identical phoneme set for all speakers available in each language. However, we separated speakers of American and British English based on speaker nationality.

The English phoneme labels, including the initial segmentation for the data, were automatically generated from the word transcriptions and speech data using the Unisyn lexicon [16] and Festival's Multisyn Build modules [17]. In the Unisyn lexicon, general American (GAM) and RP phonesets were used for American and British speakers, respectively. 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.

The Finnish and Mandarin labels were also automatically generated from the word transcriptions and speech data using an extended LC-STAR lexica [18] and Nokia's in-house TTS modules. The modules also identified utterance-medial pauses and they were added to the labels. We used phonemes instead of typical Mandarin units, initial/final [19] since we found that the phoneme-based systems perform better when the amount of adaptation data available is limited because of the smaller number of units they specify [20], [21].

The Japanese phoneme labels were also automatically generated from the word transcriptions and speech data using ATR's XIMERA TTS modules [22]–[24]. The modules also identified utterance-medial pauses and they were added to the labels.

The Spanish labels were automatically generated using new front-end modules [25] that considers the word transcriptions and several handwritten rules available in Festival modules originally developed for a Castilian Spanish diphone synthesizer called "el_diphone."

English, Spanish, and Japanese phonesets are based on IPA and Finnish and Mandarin phonesets are based on SAMPA-C. The numbers of phonemes (including the utterance-medial pauses and silences) for each language are 57, 53, 31, 26, 51, and 42 for U.S. English, U.K. English, Spanish, Finnish, Mandarin, and Japanese, respectively.

F. Phonetic, Prosodic, and Linguistic Contexts

Compared to the phonetic contexts used for ASR (e.g., preceding and succeeding phonemes), the contexts used for TTS are very rich and include various prosodic and linguistic information as well as phonetic information. The contexts we employ can be summarized in Table IV. English, Spanish, and Finnish contexts that we employ have the same structure and they contain phonetic, segment-level, syllable-level, word-level, and utterance-level features. Specifically, this includes lexical stress, neighboring phones, part-of-speech, position in syllable, etc. (see [26] for more details). In addition to these features, Mandarin contexts that we employ have tonal information. The structure of Japanese contexts are borrowed from the XIMERA TTS system and thus they are different from those for other languages: it contains phonetic, mora-level [15], morpheme, accentual, breath-group-level, and utterance-level features.

Questions used for clustering of the acoustic HMMs [27], mentioned in Section II-G, were also automatically generated. For instance the phonetic questions are automatically defined based on combinations of vowel phonetic categories such as vowel height or frontness and consonant categories such as place or manner of articulation.

G. Framework of Speaker-Adaptive HMM-Based Speech Synthesis Systems

All TTS voices are built using the framework from the "HTS-2007/2008" system [5], [28], which was a speaker-adaptive system entered for the Blizzard Challenge 2007 [29] and 2008 [30]. The HMM-based speech synthesis system, outlined in Fig. 2, consists of four main components: speech analysis, average voice training, speaker adaptation, and speech generation.

In the speech analysis part, three kinds of parameters for the STRAIGHT (Speech Transformation and Representation by Adaptive Interpolation of weiGHTed spectrogram [31]) mel-cepstral vocoder with mixed excitation (i.e., the mel-cepstrum, $\log F_0$ and a set of band-limited aperiodicity measures) are extracted as feature vectors for HMMs. These are the same features as mentioned in [32]. In the average voice training part,

TABLE IV
NUMBER OF CONTEXTS USED IN EACH LANGUAGE. ENGLISH, SPANISH, AND FINNISH CONTEXTS THAT WE EMPLOY HAVE THE SAME STRUCTURE. MANDARIN CONTEXTS THAT WE EMPLOY HAVE TONAL INFORMATION ADDITIONALLY. THE STRUCTURE OF JAPANESE CONTEXTS ARE BORROWED FROM THE XIMERA TTS SYSTEM

Contexts	English	Spanish	Finnish	Mandarin	Japanese
Phonetic	5 (quinphone)	5	5	5	5
Segmental	2	2	2	2	0
Mora [15]	0	0	0	0	3
Morpheme	0	0	0	0	12
Syllable (inc. stress and pitch accent)	22	22	22	22	12
Word	12	12	12	12	0
Tone	0	0	0	3	0
Phrase/Breath group	9	9	9	9	12
Utterance	3	3	3	3	3
Total	53	53	53	56	47

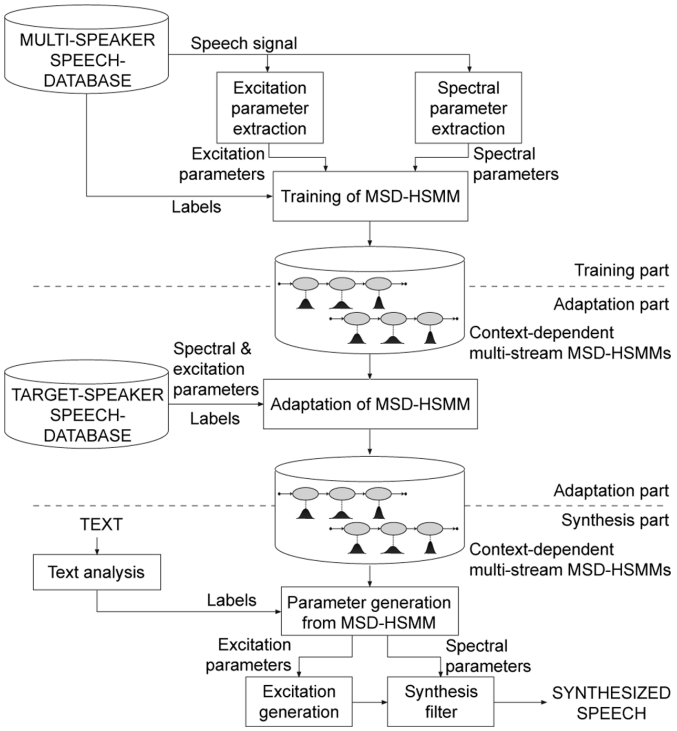


Fig. 2. Overview of the HTS-2007/2008 speech synthesis system, which consists of four main components: speech analysis, average voice training, speaker adaptation, and speech generation.

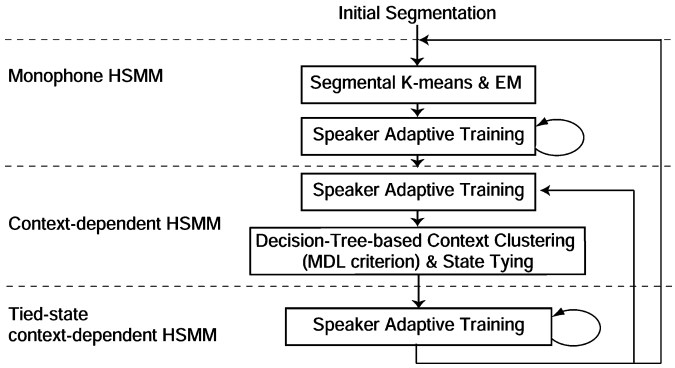


Fig. 3. Overview of the training stages for average voice models.

context-dependent multi-stream left-to-right multi-space distribution (MSD) hidden semi-Markov models (HSMMs) [33] are trained on multi-speaker databases in order to simultaneously model the acoustic features and duration. A set of model parameters (mean vectors and diagonal covariance matrices of Gaussian pdfs) for the speaker-independent MSD-HSMMs is estimated using the EM algorithm [34].

An overview of the training stages for the average voice models is shown in Fig. 3. First, speaker-independent monophone MSD-HSMMs are trained from an initial segmentation, converted into context-dependent MSD-HSMMs, and reestimated. Then, decision-tree-based context clustering with the MDL criterion [35] is applied to the HSMMs and the model parameters of the HSMMs are tied at leaf nodes. The clustered HSMMs are reestimated again. The clustering processes are repeated twice and the whole process is further repeated three times using segmentation labels refined with the trained models in a bootstrap manner [36].[3] All reestimation and resegmentation processes utilize speaker-adaptive training (SAT) [37] based on constrained maximum-likelihood linear regression (CMLLR) [38].

In the speaker adaptation part the speaker-independent MSD-HSMMs are transformed by using CMLLR or constrained structural maximum *a posteriori* linear regression (CSMAPLR) [39]. In the speech generation part, acoustic feature parameters are generated from the adapted MSD-HSMMs using a parameter generation algorithm that considers both the global variance of a trajectory to be generated and trajectory likelihood [40]. Finally an excitation signal is generated using mixed excitation (pulse plus band-filtered noise components) and pitch-synchronous overlap and add (PSOLA) [41]. This signal is used to excite a mel-logarithmic spectrum approximation (MLSA) filter [42] corresponding to the STRAIGHT mel-cepstral coefficients to generate the speech waveform.

H. Analysis of ASR Corpora—Phonetic Coverage

One clear advantage of the ASR corpora is phonetic coverage. Triphone and context coverage is a simple way to measure the phonetic coverage of a corpus. Table V shows the total number of different triphone and context types in the English corpora. Since the predefined official training data set (known as SI-284)

[3]Although we could utilize the HSMMs themselves for re-segmentation by using weighted finite-state transducers, in this case for efficiency we simply reduced the HSMMs to normal HMMs and used these to perform the Viterbi alignment.

TABLE V
PHONETIC COVERAGE OF ENGLISH MULTISPEAKER TTS AND ASR CORPORA

Corpus	Subset	Size [h]	Speakers	Triphones/speaker	Triphones/corpus	Contexts/corpus
TTS corpora						
CMU-ARCTIC	(total)	6	6	10041	10708	91247
CSTR	n/a	41	15	11462	42860	1157755
ASR corpora						
RM	(ind_total)	5	160	1091	7162	114945
WSJ0	(short/SI-84)	15	84	3287	18577	421476
WSJCAM0	(total)	22	140	3036	23534	675266
WSJ0+WSJ1	(short/SI-284)	81	284	4220	23776	1246728

TABLE VI
PHONETIC COVERAGE OF THE MANDARIN SPEECON CORPUS

Environment	Size [h]	Triphones/corpus	Contexts/corpus
Office	12	4999	71863
All environments	34	5865	181338

TABLE VII
DETAILS OF ENGLISH AND MANDARIN TTS CORPORA USED FOR BUILDING HMM-BASED TTS SYSTEMS

Corpus	Language	Speakers	Sentences
CMU-ARCTIC (BDL)	English	1	1130
CMU-ARCTIC (total)	English	6	6780
BL2009 (arctic)	English	1	1130
BL2009 (total)	English	1	9509
BL2009	Mandarin	1	6000
CSTR	English	15	29552

for WSJ1 includes WSJ0 as a part of training data, we followed instructions for the November 93 CSR evaluations and calculated them together. The predefined official training data set for WSJ0 is known as SI-84. 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 coverage of the CMU-ARCTIC speech database which includes four male and two female speakers is also shown. Details of the CSTR database are given in the next subsection.

We can see that the coverage of the complete WSJ0, WSJ1, and WSJCAM corpora is much higher than CMU-ARCTIC. This is because all speakers in CMU-ARCTIC read the same set of sentences and thus the total coverage across all speakers in the database is about the same as that of an individual speaker. 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 or used in limited domain as we do for ASR.[4]

Table VI shows the total number of different triphone and context types in the Mandarin SPEECON database. The total numbers for the office environment and the mixed environments are shown. We see the data set for the mixed environments has a much larger coverage than that of the office environment. There is a tradeoff between consistency of recording conditions and phonetic coverage.

I. TTS Speech Databases to be Compared

To enable comparison of TTS databases with the ASR corpora mentioned above, we used CMU-ARCTIC, the 2009 British English and Mandarin databases (which we refer to as "BL2009" database), and a CSTR's in-house database (which we refer to as "CSTR" database). All of these are standard purpose-built high-quality TTS databases and have very clean speech data. Details are given in Table VII.

The CMU-ARCTIC database has six speakers, each of whom reads the same set of 1130 phonetically balanced sentences, corresponding to about 1 hour of speech data per speaker. In Section V-C, an American male speaker from the database, "BDL," was chosen as one of out target speakers and his speech data was utilized for speaker adaptation of the average voice models and for training of speaker-dependent models.

The British and Mandarin BL 2009 corpora were released for the 2009 Blizzard Challenge [43] and they have a male English RP speaker and a female Mandarin speaker. They include 9509 and 6780 sentences corresponding to 15 hours and 10.5 of speech, respectively. The English BL 2009 corpus has the ARCTIC sentences above as one of subsets. In Sections V-C and D, they were also chosen as target speakers for speaker adaptation.

The above TTS corpora were mainly used only for speaker adaptation or for training speaker-dependent models. For the training of the average voice models which provide the starting point for speaker adaptation, we used the CSTR database having 41 hours of very clean speech data uttered by 15 speakers and compared it with several English ASR corpora in Section V-C. In total, the CSTR database includes 29 552 sentences. The original HTS-2008 system that used the CSTR database for the training of the average voice models was evaluated in the 2008 challenge [30]. Since in the CSTR database speakers utilize different sets of texts, its context coverage is as high as that of the ASR corpora. As shown in Table V, it has about four times as many triphones as the CMU-ARCTIC database and 1.4 times as many as the SI-284 sets. In fact, the system had the equal best naturalness and the equal best intelligibility on the Arctic data in the 2008 challenge [28]. The system was also found to be as intelligible as human speech [28]. Thus, we considered the very

[4]Although the context coverage including prosody contexts of the RM corpus is slightly better than CMU-ARCTIC, its triphone coverage is critically worse than CMU-ARCTIC.

TABLE VIII
NUMBER OF LEAF NODES AND FOOTPRINTS FOR SPEAKER-DEPENDENT (SD) AND SPEAKER-ADAPTIVE SYSTEMS. NOTE THAT SYSTEMS THAT WERE NOT EVALUATED IN THE LISTENING TESTS MENTIONED LATER ARE ALSO INCLUDED. (a) ENGLISH SYSTEMS. (b) MANDARIN SYSTEMS. (c) OTHER SYSTEMS

(a)

			Number of leaves in decision tree				Footprint [Mega-Bytes] Acoustic models		Linear transforms
Corpus	Subset	Size [h]	Mel-cepstrum	log F_0	Aperiodicity	Duration	HTK	hts_engine	HTK
SD models									
CMU-ARCTIC	(BDL)	1	871	2118	705	658	50	1.5	n/a
BL2009	(total)	15	5833	27137	6790	4045	517	13.0	n/a
Average voice models									
CMU-ARCTIC	(total)	6	2311	11613	1504	3194	118	3.1	0.5
CSTR	n/a	41	9380	49269	5138	8539	1592	31.5	0.7
RM	(ind_total)	5	2122	12417	2839	3733	334	5.8	0.1
WSJ0	(short/SI-84)	15	2945	26952	2624	13165	669	11.0	0.5
WSJCAM0	(total)	22	3599	40326	3237	23641	981	13.0	0.2
WSJ0+WSJ1	(short/SI-284)	88	10861	105940	9202	51567	1697	34.0	0.9

(b)

			Number of leaves in decision tree				Footprint [Mega-Bytes] Acoustic models		Linear transforms
Corpus	Environment	Size [h]	Mel-cepstrum	log F_0	Aperiodicity	Duration	HTK	hts_engine	HTK
SD models									
BL2009	n/a	11	4837	14175	4654	3335	400	9.5	n/a
Average voice models									
SPEECON	Office	12	2373	15320	3238	3474	442	7.0	0.1
SPEECON	All environments	34	6272	33905	6378	7695	681	17.0	0.3

(c)

			Number of leaves in decision tree				Footprint [Mega-Bytes] Acoustic models		Linear transforms
Corpus	Language	Size [h]	Mel-cepstrum	log F_0	Aperiodicity	Duration	HTK	hts_engine	HTK
Average voice models									
SPEECON (Office)	Finnish	12	1383	16682	2325	6950	279	5.3	0.1
JNAS	Japanese	20	2388	34642	2705	9995	527	9.2	0.3
GlobalPhone	Spanish	22	2925	52088	2308	27193	1377	17.0	0.5

clean and also contextually rich database as an ideal case for other ASR corpora.

J. Number of Leaves in Decision Trees

Table VIII shows the number of leaves of each of the decision trees for each system built on the speaker-independent subset of each multispeaker corpus. For comparison, the table shows the number of leaves for speaker-dependent HMMs trained on the BDL subset of the CMU-ARCTIC, the British English and the Mandarin BL 2009 corpora. Systems that were not evaluated in the listening tests mentioned later are also included in this table for reference.

From the tables, we see that the trees for mel-cepstral and aperiodicity elements of voices built on WSJ0 and WSJCAM0 have fewer leaves than those of English speaker-dependent (SD) HMMs trained on the English BL2009 corpus, although the WSJ0 and WSJCAM0 databases are similar in size to those for the SD-HMMs. On the other hand, they have almost the same number or more leaves for $\log F_0$ and duration parts. Trees for the voice built on the office subset of the Mandarin SPEECON database have similar sizes to those for Mandarin SD-HMMs and we can see the same tendency also for the Mandarin systems. The fact that they include various dialectal accents may partially explain the greater number of $\log F_0$ leaves. However, further investigation is required, especially in the case of the much greater number of duration leaves of the English systems trained on the WSJ databases. It can be seen that the Mandarin system using data from all environments has more leaves than the SD-HMM system or the one using data from office environments only.

K. Footprint of Each System Built

Table VIII also shows the footprints of speaker-independent HMMs and linear transforms for each system built. For reference the footprints of speaker-dependent HMMs are also shown. Since we used a single Gaussian for each leaf node, the number of leaves is a dominant factor for the footprint of the acoustic models. It shows the footprints in both the standard HTK format and the hts_engine format that maintains only statistics required for use by synthesis modules.

We can also see that the speaker-adaptive HMM-based speech synthesis is efficient in terms of footprint. For example, the footprint of the average voice model (which is speaker independent

TABLE IX
NUMBER OF THE TTS VOICES BUILT ON ASR CORPORA

Language	Corpus	Subset	TTS voices
English	Total		651
	WSJ0	short	84
	WSJ0	long	12
	WSJ0	very long	3
	WSJ0	dev, eval	18
	WSJ1	short	200
	WSJ1	long	25
	WSJCAM0	total	140
	RM	ind_total	160
	RM	dev_total	12
Finnish	SPEECON	office	200
Mandarin	SPEECON	office, all	500
Japanese	JNAS	(random)	100
Spanish	GlobalPhone	all	100
Total			**1554**

and thus can be shared among many speakers) trained on WSJ0 and WSJCAM0 is as compact as that of the SD-HMMs trained on the English BL2009 corpus in the hts_engine format. Furthermore, the speaker-dependent footprint, a set of linear transforms for each speaker, is less than 1 MB and thus we can increase the number of voices efficiently. The average voice model trained on the SI-284 set has larger footprint than that of the SD-HMMs in the hts_engine format. However, sharing the large average voice models among hundreds of speakers leads to more efficient footprint overall than maintaining hundreds of separate SD-HMMs.

L. Demonstration of the TTS Voices

Since we aim to give a fair impression of the quality of synthetic speech built from each corpus and to discuss the usefulness of the ASR corpora, we followed predefined training recipes for each corpus, built speaker-adaptive gender-independent HMM-based TTS systems from major subsets of each corpus separately, and adapted them to all speakers available. A summary of the number of TTS voices built from each ASR corpus is given in Table IX. Audio samples are available from http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/ASRcorpora.html

Careful listening reveals 1) that the quality of synthetic speech varies according to which corpus is used to train the average voice models, or by the amount of adaptation data used and 2) that there are a few speakers whose synthetic speech sounds worse than that of other speakers who have the same amount of adaptation data from within the same corpus.

For the first case, our previous analysis has already shown that the amount of adaptation data required for reproducing speaker similarity above a certain level varies by target speakers (and acoustic features) and ranges from three minutes to six minutes in terms of speech duration [44] and also that the naturalness of the synthetic speech generated from the adapted models is closely correlated with the amount of data used for training the average voice models [39].[5]

This directly explains the relatively low quality of voices built on the RM corpus since the small corpus does not satisfy the two conditions above: the total duration of training data for the average voice model is just five hours and the duration of adaptation data in the RM corpus averages 3.8 minutes. Evaluation results reported in Section V-C also highlighted this issue. The first condition also explains unstable adaptation performance of voices built on the SPEECON databases, since the duration of adaptation data in the SPEECON databases averages 3.8 minutes and that is not always enough for all the speakers.

The interesting phenomenon observed in the second case is new and analogous to the familiar situation in ASR, where WER varies widely across some speakers and is especially high for a small number of speakers [47]. This is investigated in Section IV-F.

M. Geographical Representation and Online Demo

One of important advantages of using the ASR corpora is the large number of speakers as we can see in Table IX. Building TTS voices on such data allows the creation of many more voices than has previously been possible for TTS.

In fact, we believe *this is the largest known collection of synthetic voices in existence*. We built so many voices (1500+ voices built on ASR corpora plus several voices built on TTS corpora using the same techniques) that it became impossible to represent them in list or table form. Instead, we devised an interactive geographical representation, shown in Fig. 4.

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. Then right box shows list of speakers that user can choose with speakers' gender and nationality. This is based on Google Maps and AJAX Language (Translation) APIs[6] as well as our Festival TTS system running on a University of Edinburgh server.

This geographical representation, which includes an interactive TTS demonstration of many of the voices, is available from http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/map.html. Clicking on a marker will play synthetic speech from that speaker, as shown in Fig. 5. Currently, the interactive mode supports all English and some of the Spanish voices. For other languages only presynthesised examples are available, but we plan

[5]We also know that gender-dependent average voice models provide better speaker adaptation performance than gender-independent average voice models for TTS [39]. Please bear in mind, however, that the purpose of this demonstration is to give benchmarks that can be easily related to the field of ASR. For example, the predefined training sets such as SI-84 or SI-284 are the de facto standard for training clean acoustic models for ASR, and HTK provides benchmark scores on the speaker independent set of the RM corpus. Obviously, we may use HMMs built for TTS purposes on ASR corpora as acoustic models for ASR [45], [46] and thus we can easily compare ASR scores of TTS HMMs with scores reported in ASR literature. On the other hand, if we explore the best quality of synthetic speech in the ASR corpora, we should combine these ASR corpora and train larger and gender-dependent average voice models as we can guess from the previous analysis results.

[6]http://code.google.com/intl/ja/apis/ajaxlanguage/

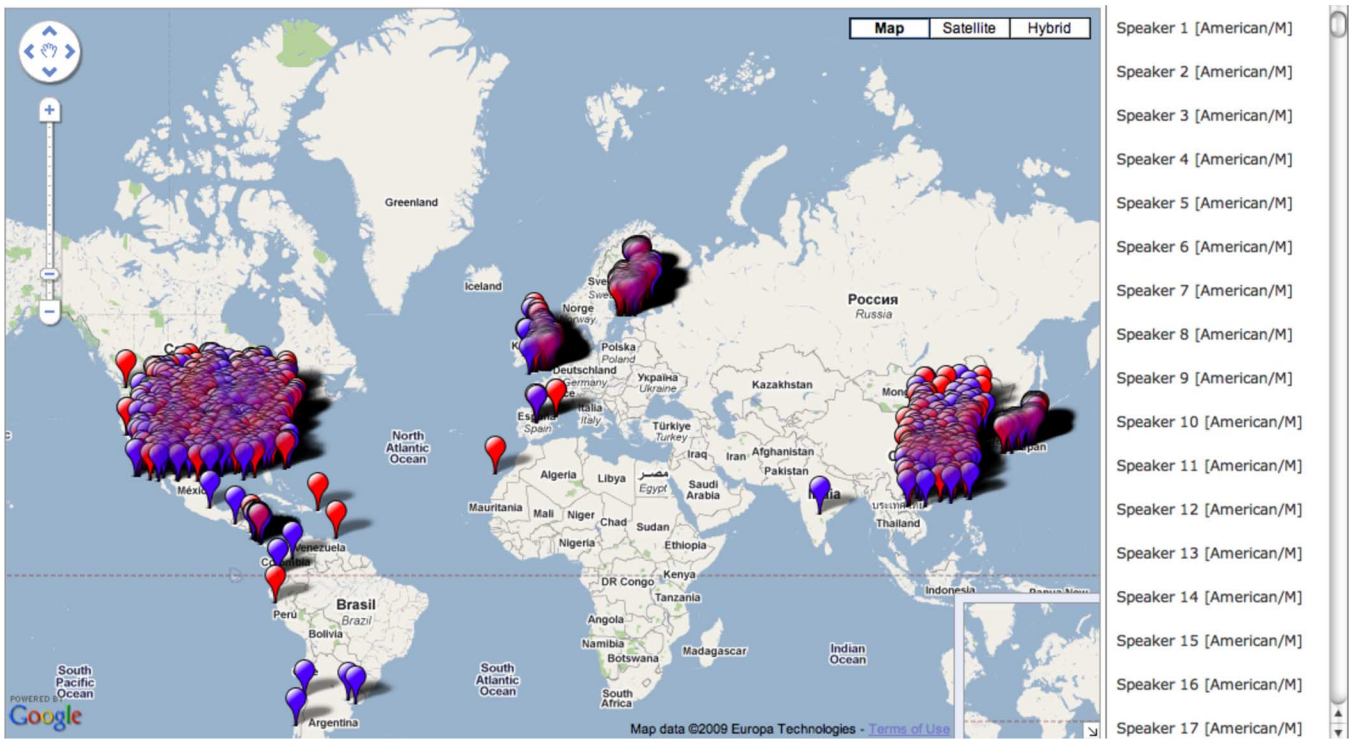


Fig. 4. Geographical representation of TTS voices trained on ASR corpora used for EMIME projects. Blue/dark gray markers show male speakers and red/light gray markers show female speakers. Available online via http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/map.html.

to add an interactive text-to-speech feature in the very near future.

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 the EMIME project,[7] whose goal is personalized cross-lingual speech-to-speech translation systems. For example, if a user's mobile device for the speech-to-speech translation systems has GPS functions, we would be able to automatically choose and utilize appropriate voices based on that user's location, obtained from GPS. Furthermore if desired we may perform cross-lingual speaker adaptation on the chosen models [48], [49].

III. Evaluation Methodologies

For the evaluation of synthetic speech, both objective measures and formal listening tests have been used. This section explains the details of the evaluation methodologies.

A. Objective Measures

Voices for a large number of speakers can be built during the training of the speaker adaptive HMM-based synthesizer. However, in some cases, there were too many speakers to evaluate by formal listening tests. The listening tests were therefore principally used for only a few target speakers. Objective measures, on the other hand, could be used for many or all speakers. Here it is important to recognize that these objective measures do not perfectly measure the quality of synthetic speech. They generally only weakly correlate with perceptual scores obtained from listening tests [50], [51].

[7]The FP7 Effective Multilingual Interaction in Mobile Environments (EMIME) project http://www.emime.org

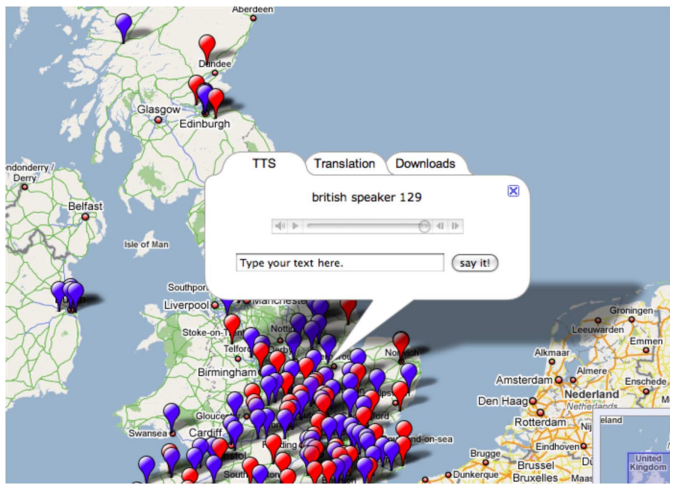


Fig. 5. All English and some of the Spanish HTS voices can be used as online TTS on the geographical map. For other languages only presynthesized examples are available.

For the calculation of all the objective measures, the synthetic speech and natural speech must be aligned frame-by-frame. In order to do this, the synthesis model for the test sentence is force-aligned with the natural speech. From this alignment, the phoneme durations of the natural speech are obtained. The model is then used to generate synthetic speech with exactly those phoneme durations (within phonemes, the usual duration model is used to obtain the state durations [52]) so that the synthetic and natural utterances have exactly the same durations

and thus a one-to-one correspondence between their frames can be used to calculate the objective measures.

B. Mel-Cepstral Distance

To measure the accuracy of the spectral envelope of the synthetic speech, we use "average mel-cepstral distance" (MCD) which is a popular objective measure used in speech coding or parametric speech synthesis (e.g., [39], [53]). When the analysis order of mel-cepstral analysis is high enough, Parseval's theorem means that the mel-cepstral distance can be viewed as an approximate log spectral distance between the synthetic speech and natural speech.

After silence and pause regions are ignored, the Euclidean distance between the mel-cepstral parameters of the natural and synthetic examples is computed.

C. Root-Mean-Square-Error of $\log F_0$

To measure the accuracy of the F_0 contour generated by the model, the second objective measure we calculate is the root-mean-square-error (RMSE) of $\log F_0$. Since F_0 is not observed in unvoiced regions, the RMSE of $\log F_0$ is calculated only using regions where both the generated and the actual speech are voiced.

The RMSE values of $\log F_0$ are shown in "cent." The cent is a logarithmic unit of measure used for musical intervals and musical scales. 1200 cents are equal to one octave (a frequency ratio of 2:1). An equally tempered semitone (the interval between two adjacent piano keys) is 100 cents.

D. Reference Materials

Natural speech utterances are required for all these objective measures, since the objective measures are computed between acoustic parameters generated from HMMs and ones extracted from the natural speech utterances. Test sentences included in neither the training nor the adaptation data were used for the objective evaluation.

E. Listening Test Design

The key properties of the synthetic speech that must be evaluated are: naturalness, intelligibility, and similarity to the original speaker. Therefore, and because the web-based listening test infrastructure was already in place, we adopted a design based on that of the Blizzard Challenge 2008 [29], [30]. We used the software developed for that Challenge which comprises: a web-based listening test that runs in any standard browser; Perl CGI scripts running on a University of Edinburgh server; results stored in a MySQL database; R scripts to compute statistics and produce graphical output.

To evaluate naturalness and similarity to the target speaker, 5-point mean opinion score (MOS) and comparison category rating (CCR) tests are used. The scale for the MOS test runs from 5 for "completely natural" to 1 for "completely unnatural." The scale for the CCR test runs from 5 for "sounds like exactly the same person" to 1 for "sounds like a totally different person" and a few example natural sentences from the target speaker are provided as a reference.

To evaluate intelligibility, the subjects are asked to transcribe semantically unpredictable sentences by typing in the sentence they heard; the average word error rate (WER) is calculated from these transcripts (an automatic procedure is used, which corrects spelling mistakes and typographical errors). The evaluations are conducted via a standard web browser.

F. Format Used to Report Results

We used the same conventions as the Blizzard Challenge for reporting results [54]: "Standard boxplots are presented for the ordinal data where the median is represented by a solid bar across a box showing the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers beyond this are represented as circles. Bar charts are presented for the word error rate interval data." In addition average scores are marked as "$\times$" within the boxplots.

The differences in the results for all three sections are measured by the same test used in the Blizzard Challenge: a Wilcoxon signed rank test with $\alpha = 0.01$ and Bonferroni correction.

G. Listeners

Different sets of listeners were collected for individual listening tests. They are natives or non-natives with sufficient understanding of the target languages. Furthermore one of the listening tests was performed via the 2009 Blizzard Challenge. Thus, we cannot directly compare results across each listening test.

H. Scenarios

There are three scenarios depending on whether the ASR corpora are used for either (or both) training of the average voice models or speaker adaptation. We have evaluated the use of the "found" audio [5], in a context very similar to one of the present scenarios, where the average voice models are trained on the TTS corpora and adaptation data is chosen from the ASR corpora.

Hence, this paper focuses on the other two scenarios: the case where the ASR corpora are used for training the average voice models and the adaptation data is chosen from the purpose-built TTS corpora, and the most difficult case where the ASR corpora are used both for training the average voice models and for speaker adaptation. The former case is reported in Section V and the latter in the section which now follows.

IV. Evaluation for HMM-Based TTS Systems Constructed From ASR Corpora

In this section, we analyze the performance of speaker-dependent (SD) and speaker-adaptive (SA) HMM-based TTS systems constructed from the ASR corpora only and assess how different their tendencies are from our previous analysis [5], [39], where performance analysis of the systems constructed from TTS corpora have been reported. We also analyze speaker distributions and their correlations to the quality of synthetic speech generated from adapted HMMs.

A. Average Voice Model Training Data

Since we had used the equal amount of data for each training speaker of the average voice models in the past analysis and since the amount of data available for each speaker in the ASR

corpora may vary by subsets and even recording sites, we built two kinds of gender-dependent average voice model from short term, long term (excluding the speakers from very long term), development, and evaluation subsets of the WSJ0 corpus using different training data sets. 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 10 847 sentences for male average voice models in conditions 1 and 2, respectively, and for female average voice models there are 3000 and 12 151 sentences, respectively. They have 5.7 hours, 21.1 hours, 5.9 hours, and 24.6 hours of speech, 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.

B. Speaker-Dependent Model Training Data

In general, training of speaker-dependent models requires $O(10^3)$ utterances and only the very long subset of the WSJ0 corpus is available for the models. Since this subset has only two males and a female, we simply chose a male speaker 001 and a female speaker 002 as target speakers for listening tests in this section.

In order 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 two target speakers included in very long subset. These sentences are also used as the adaptation data for the two average voice models mentioned in previous section.

C. Objective Evaluation of SD and SA Systems

Table X 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, 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 relatively consistent with previous results except the SD and SA systems using 2 hours of speech data were still comparable to TTS databases.

On the other hand, the RMSE of $\log F_0$ for the speaker 002 shows different tendencies. All the systems using 2 hours of target speaker speech data have worse RMSE than those using 1 hour of data although the absolute values are better than those for the speaker 001.

One of possible explanations 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 is a natural consequence of using ASR data since the speakers are not trained voice talents. Although usually speaker adaptation is used for less target speaker data, more sophisticated strategies are required to cope with such changeable characteristics. For example, if the adaptation data includes acoustic fluctuation that should not be explained by linguistic contexts such as speaker's mood or fatigue, preselection or preclustering of adaptation data should be added to the adaptation process.

TABLE X
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. (b) 1 HOUR OF TARGET SPEAKER DATA. (c) 2 HOURS OF TARGET SPEAKER DATA

(a)

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

(b)

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

(c)

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

D. Subjective Evaluation of SD and SA Systems

We chose the male speaker 001 as the target speaker for the subjective (listening test) evaluation. 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 [30] for details of the Latin Square experimental design.) The number of listeners for this experiment was 26.

Fig. 6 shows the results. The perceptual evaluation reveals the same tendencies as the objective evaluations. The speaker-adaptive systems using 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

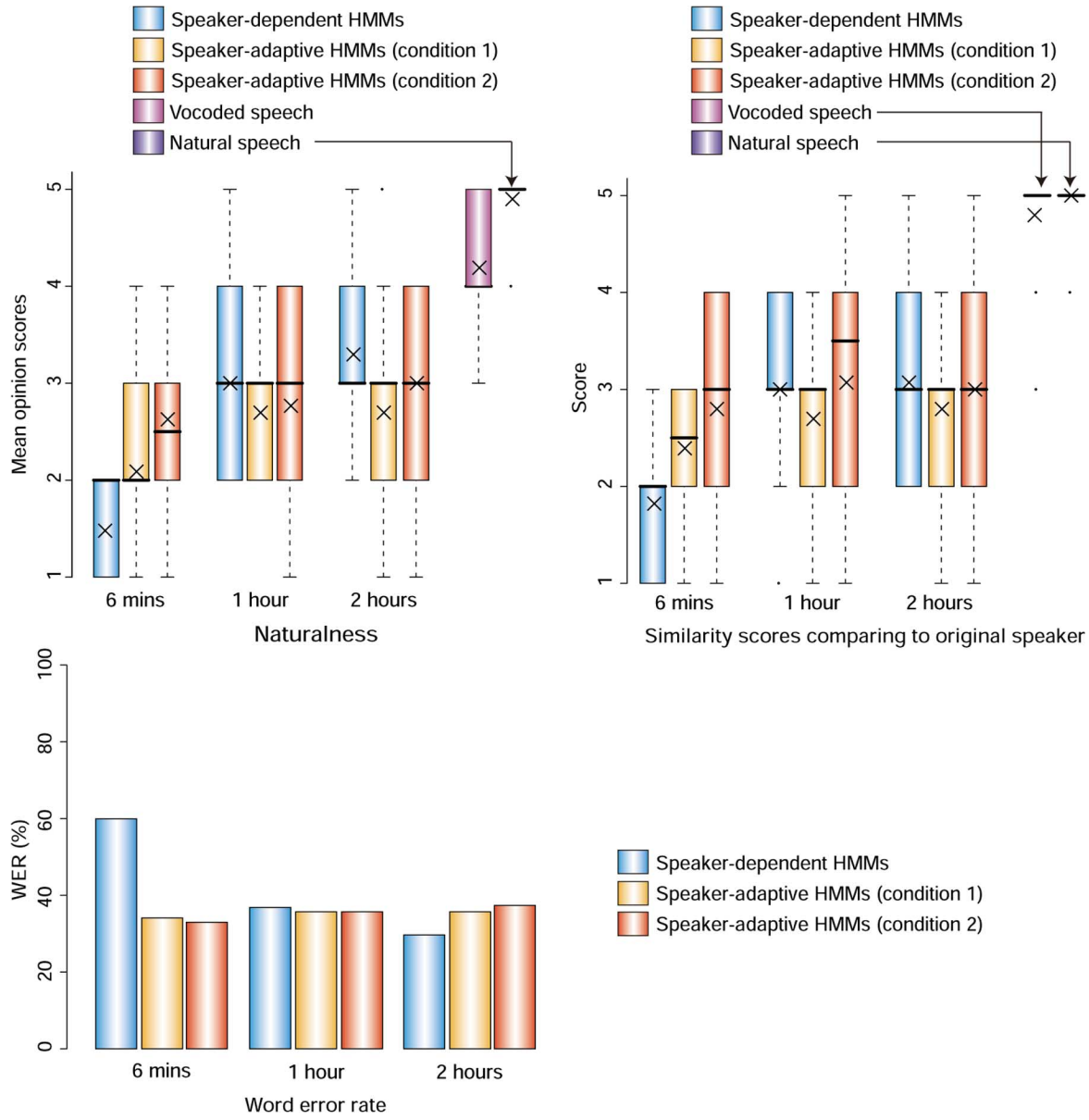


Fig. 6. Subjective evaluation results for speaker-dependent and speaker-adaptive HMM-based TTS systems built on ASR corpora. The target speaker used is a male speaker "001" included in very long subset of the WSJ0 corpus.

is less than about 1 hour, speaker-adaptive systems outperform speaker-dependent systems in every way ($p < 0.01$). 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 than the speaker-adaptive system. In the intelligibility test, only the speaker-dependent system using six minutes of speech data was found to be significantly worse the other systems ($p < 0.01$). Other differences between WERs were not statistically significant.

We note that previous work on TTS databases has indicated that 100 utterances (approximately 6 minutes) of adaptation data are enough to adapt an average voice to the characteristics of a target speaker [44]. On the other hand, this figure shows that 15 minutes of data achieve a median opinion score of only 2.5 for naturalness. The previous work also indicated that the SD and SA systems using 2 hours of speech data were comparable, whereas the SD systems start to become better than SA systems on this ASR database. We attribute this low score to the noisiness of the adaptation data and conclude that more target speaker data were needed to obtain a reasonable naturalness rating.

E. Multidimensional Scaling of 120 HTS Voices Adapted and Average Voices

Rather than visualizing speakers by placing them in a geographical space, we can place them in a space derived from the properties of the speech and can analyze speaker distributions. There are several conventional approaches to visualize speakers or speaking style based on acoustic models or acoustic features [55], [56].

A similar visualization can be straightforwardly achieved using the HTS voices built and multidimensional scaling (MDS) [57]. Using all test sentences from the Blizzard Challenge 2008,

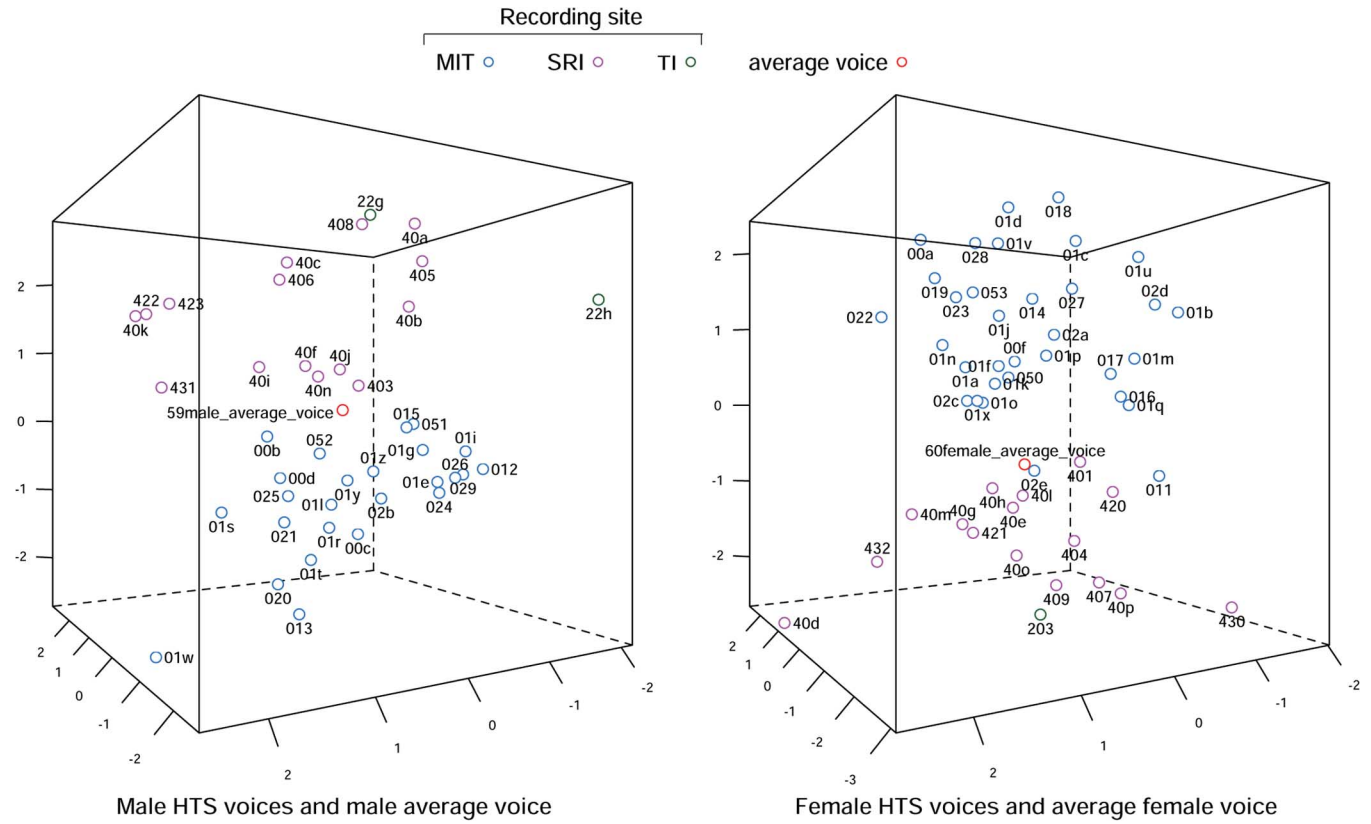


Fig. 7. Multidimensional scaling of 120 HTS voices trained on the WSJ0 corpus. The three characters at each point correspond to the name of each speaker in the database. Left part shows the male speakers and male average voice and right parts shows the female speakers and female average voice. A demonstration movie for two-dimensional MDS representation is available via http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/mov/HundredsHTS.mov.

we generated a set of speech samples from the gender-dependent average voice models and all the HTS voices that equally had a hundred of adaptation sentences. For the average voice models, we used "condition 2" of the previous evaluation. We then calculated the average mel-cepstral distance between the speech for all pairs of voices, placing the values in mel-cepstral distance tables. 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 the same. Then we applied a classic multidimensional scaling technique [57] to the mel-cepstral distance table and examined the resulting three-dimensional space, which is shown in Fig. 7. On the left-hand side of the figure, the MDS of the male speakers and male average voice appear and on the right, that of the female speakers and female average voice.

The axes of this space do not have any *predefined* 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, in the MDS for male speakers, speakers 012, 01i, 01e, 026, and 029 are similar to one another (in terms of mel-cepstral distance) and speakers 22h, 422, 423, 40k are relatively different from other speakers. If we need to analyze performance of "outlier" speakers it would be reasonable to choose speakers based on this MDS representation.

Instead of the geographical GUI above, we may use the MDS space for an alternative GUI for the HTS voices (see a provided URL in the caption for Fig. 7). More importantly, since we could only use very few target speakers in formal listening test, we should investigate the distribution or tendency of many speakers in other ways, such as MDS.

On examining the figure in detail, we noticed that all three-characters codes (corresponding to the names of speakers) distributed in the bottom part start with 0 and the codes for speakers distributed in top part start with 4. The first character of the names represents recording site for these speakers (0: MIT, 4:SRI, and 2:TI) [8]. Therefore, we assigned different colors to each recording site in the figure.

It is apparent that recording conditions were not consistent among the recording sites although the same microphones were utilized. Furthermore, acoustic differences due to the inconsistent recording conditions are greater than acoustic differences between speakers since there is an obvious border between them. Thus, the average voice models trained on these speakers are located at the center of recording conditions rather than the center of the speakers. If we use additional hierarchical transforms to normalize the recording conditions as well as speaker transforms [58], [59], it would be possible to make the average voice model more compact and more efficient.

F. Subjective Evaluations of Speakers Included in WSJ0 Corpus and Average Voices

Next we analyze fluctuation of the quality of synthetic speech generated from models adapted from the same average voice models using the same amounts of adaptation data chosen from the same corpus. For this purpose we utilized the 59 male voices

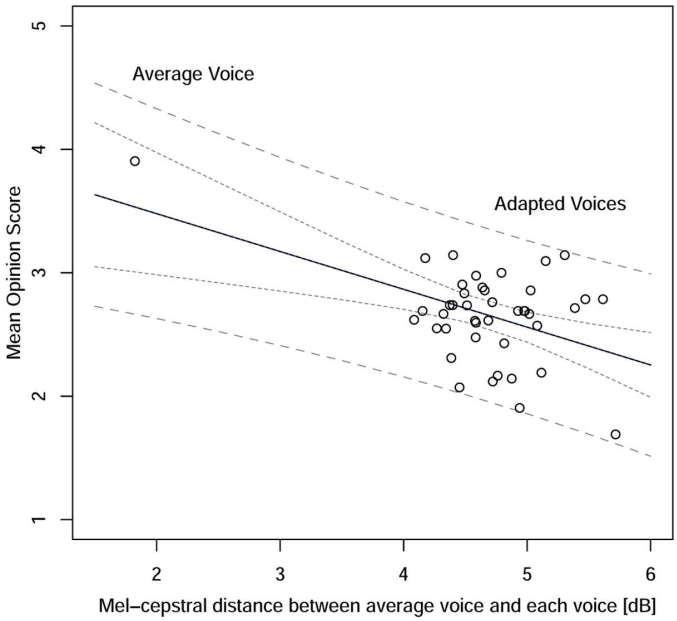


Fig. 8. Scatter plot of the mean MOS scores of 59 male voices adapted and a male average voice model. Each dot represents either the male speaker or male average voice. Horizontal axis shows the mel-cepstral distance from the average voice. Vertical axis shows the mean MOS score obtained for each voice. This also represents a linear regression function fitted and its 95% confidence and prediction intervals. For computation of the mel-cepstral distance for the average voice itself, random-sampling-based parameter generation algorithm [60] was used.

and a male average voice used for MDS in the previous section and evaluated their naturalness using the MOS test in which four test sentences were randomly chosen from all the test sentences used for MDS above. The number of listeners was 40.

The Pearson product–moment correlation coefficient between the mean MOS scores obtained in the evaluation and the first axis of MDS which represents the recording sites, the second axis, and mel-cepstral distance between average voice and each voice (which can viewed as a transformed distance of the voice) are -0.13, -0.38, and -0.48, respectively. In a word, the MOS scores obtained are not correlated with the recording sites and associated recording condition differences. However it is somewhat correlated inversely with mel-cepstral distance from the average voice. Its 95% confidence intervals are from -0.20 to -0.68.

Fig. 8 shows the scatter plot of the mean MOS scores for the voices and the mel-cepstral distance from the average voice. For computation of the mel-cepstral distance for the average voice itself, the random-sampling-based parameter generation algorithm [60] was used. This also represents a linear regression function fitted and its 95% confidence and prediction intervals.

We can see that as the mel-cepstral distance from the average voice becomes larger, the MOS scores generally become worse. Readers might also be surprised at the highest scores of the average voice in the evaluation (the mean MOS score is 3.9.). A similar tradeoff phenomenon between transformed distance and quality reduction of synthetic speech has been observed in voice conversion [61].

In addition to the transformed distance, we hypothesize that there is a psychological reason: Langlois and Roggman have shown that averaged faces look more attractive than individuals in their paper entitled "Attractive Faces are Only Average" [62]. In a similar way, a likely psychological explanation for the higher score of the average voices is that *attractive voices are also average*. This is a very interesting aspect which has a deeper meaning and implies a new direction of the statistical parametric speech synthesis approach since the statistical averaging effect, which is an acknowledged weakness of current HMM-based speech synthesizers might have the potential to produce voices that sound more attractive than individuals.

V. Evaluation for Average Voice Models Trained on ASR Corpora and Adapted on TTS Corpora

Finally, we analyze the situation most likely to be encountered in real life, where average voice models are trained on ASR corpora and are adapted to target speakers chosen from TTS corpora. In this scenario, we can use both advantages, that is, high context coverage of the ASR databases and high-quality speech of the purpose-built TTS databases. We also evaluate the effect of the quantity and inconsistent recording conditions of ASR data used for training of the average voice models together.

A. Average Voice Model Training Data

We utilize the English and Mandarin average voice models in this section. For the training data of the average voice models, we have used the predefined speaker-independent training data set for each corpora mentioned earlier. The context coverages of the data set are shown in Tables V and VI. Model complexity and footprints for each of the systems built on each of the datasets are shown in Table VIII(a) and (b).

B. Target TTS Database

For the target TTS databases from which adaptation data is chosen, we used CMU-ARCTIC, British English and Mandarin BL 2009 databases. For details see Section II-A.

C. Subjective Evaluation of the Quantity of Data Used for Training of the Average Voice Models

Using the Arctic subsets (ca. one hour of speech data) of the British speaker's corpus, we adapted the English average voice models trained on each ASR corpus having various amounts of data and compared synthetic speech generated from adapted models to see the effect of the quantity of the training data.[8]

At the same time we also evaluated the original HTS-2008 system that used the CSTR database, which is very clean, contextually rich ideal TTS database. Note that since the CSTR database comprises male speakers only and since the SI-284 sets comprises both genders equally, they include almost the same amounts of male speakers' data. For listening tests, we utilized all test sentences used for the 2007, 2008, and 2009 challenge. The number of listeners who completed the listening test was 68.

The evaluation results are shown in Fig. 9. In the MOS evaluation on naturalness, the reference system using the CSTR database was found to be significantly better than other systems ex-

[8]In reality, both supervised and unsupervised adaptation were evaluated together in the listening tests. However, there were not significant differences between them and thus we omitted results for the unsupervised versions. For full results, see [63].

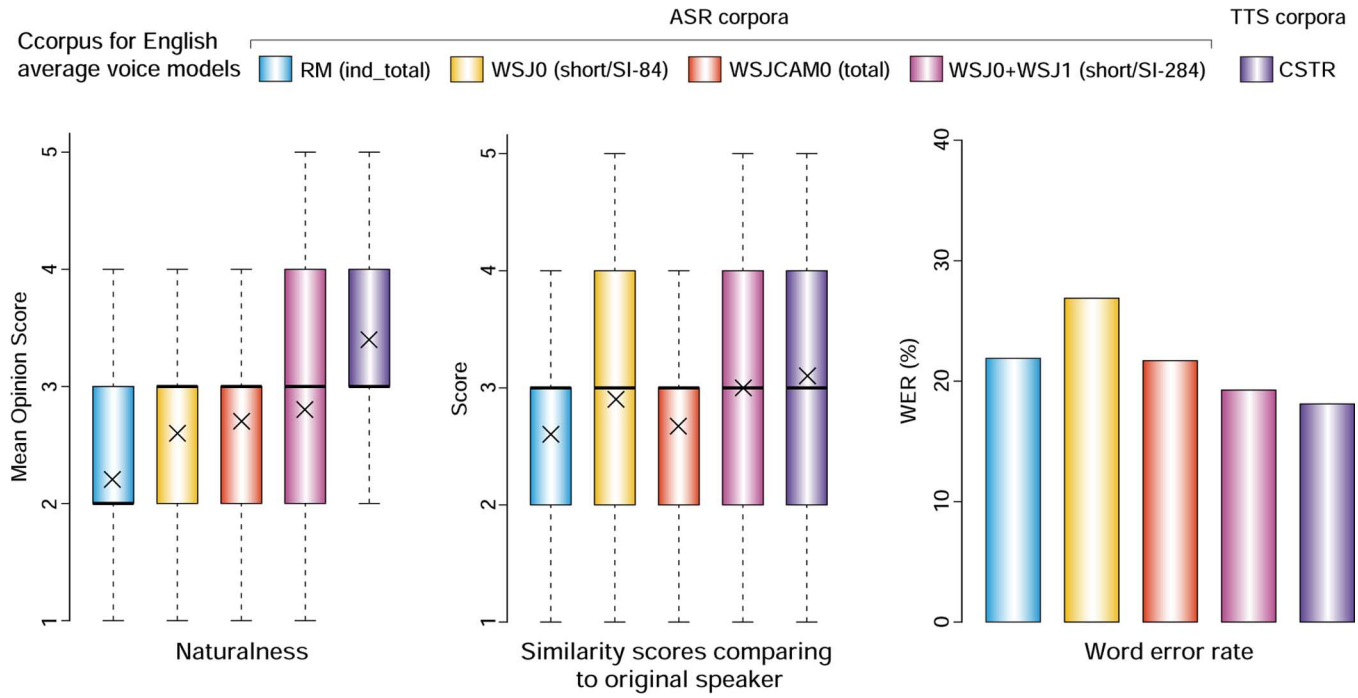


Fig. 9. Subjective evaluation results for speaker-adaptive HMM-based TTS systems built on various ASR and ideal TTS corpora. The target speaker used is a British male speaker included in the 2009 Blizzard Challenge corpus. The total amounts of speech data used for RM, WSJ0, WSJCAM0, WSJ0+WSJ1, and CSTR are 5 hours, 15 hours, 22 hours, 81 hours, and 41 hours. Note that the CSTR database includes gender-specific waveforms only.

cept the SI-284 system ($p < 0.01$). The SI-284 system was also found to be significantly better than the RM system ($p < 0.01$). Other differences in the MOS evaluation were not statistically significant. In the similarity (CCR) evaluation, there was no statistically significant difference. In the intelligibility evaluation (WER), the reference and SI-284 systems were found to significantly better than the SI-84 system ($p < 0.01$).

Overall, we can see the average voice models using larger amounts of data provide better results in general. Compared to the reference system using the CSTR database, there are several negative conditions for the SI-284 systems in addition to the ASR recording quality. For example, the SI-284 system did not have any British speakers and was gender-independent, whereas the reference system had many British speakers and gender-dependent. However, we can also see that the SI-284 system provided relatively good performance close to that of the reference system. This is a promising result since from publicly available ASR databases we can create average voice models that have good performance close to ideal TTS ones.

The original HTS-2008 system has also been found significantly better than the speaker-dependent (SD) systems in terms of similarity and naturalness on this small Arctic subset [28]. Therefore, we expect the SI-284 systems would also have better performance than the SD systems similarly. To confirm this, we performed additional listening tests using the American speaker BDL included in the CMU-ARCTIC database as a target speaker and compared the RM, SI-84, and SI-284 systems adapted to the speaker with the SD system. We used all the Arctic sentences as adaptation sentences. In the same way as previous tests, test sentences used for this listening tests were randomly chosen from all test sentences for the 2007, 2008, and 2009 challenge. The number of listeners are 40. Table XI shows the comparison results with the SD system in which we can see the SI-284 system outperforms the SD system.

TABLE XI
SUBJECTIVE EVALUATION RESULTS FOR SPEAKER-ADAPTIVE HMM-BASED TTS SYSTEMS BUILT ON VARIOUS ASR AND A SPEAKER-DEPENDENT HMM-BASED TTS SYSTEM BUILT ON CMU-ARCTIC CORPUS. TARGET SPEAKER IS BDL (AMERICAN MALE)

Corpus	Subset	Size (h)	MOS
SD models			
CMU-ARCTIC	(BDL)	1	2.5
Average voice models plus adaptation			
RM	(ind_total)	5	2.3
WSJ0	(short/SI-84)	15	2.6
WSJ0+WSJ1	(short/SI-284)	81	2.8

From these positive results, we conclude that these clean ASR corpora are useful even for the training of the average voice models used in speaker-adaptive HMM-based speech synthesis.

D. Subjective Evaluation of Mixed Recording Conditions

Since this is our first challenge using speech with background car noise, etc. for TTS, we have not used any noise suppression techniques such as Wiener filtering. Instead we simply changed acoustic features from mel-cepstra to mel-generalized cepstra [64] with cubic-root compression of amplitude and applied SAT which includes cepstral mean normalization (CMN) and cepstral variance normalization (CVN) implicitly and trained the average voice models as normal. This was motivated by *multi-style* training used in the ASR field [65]. The use of mel-generalized cepstra was also motivated by ASR. The mel-generalized cepstra are similar to PLP features in terms of spectral representation [64]. Thus, we expect that they should provide similar

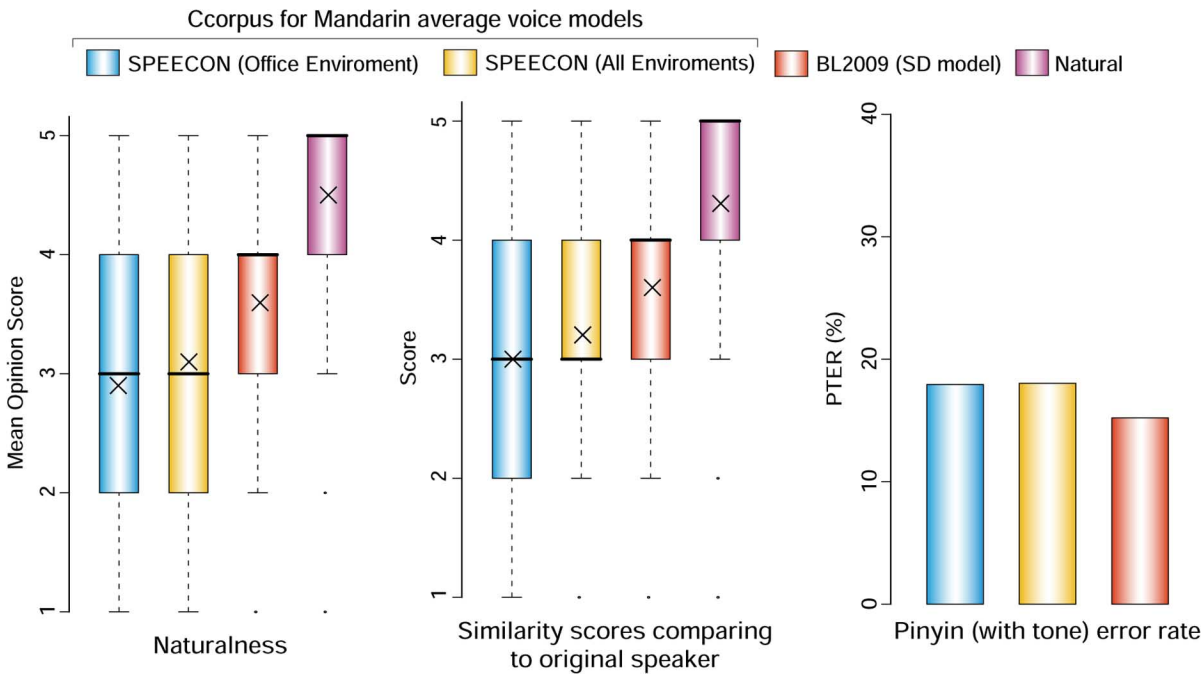


Fig. 10. Subjective evaluation results for speaker-adaptive HMM-based TTS systems built on speech data recorded in office and various conditions. For reference natural speech and speaker-dependent systems were also evaluated. The target speaker used is a Mandarin female speaker included in the 2009 Blizzard Challenge corpus.

robustness to noise as the PLP features, which are known to give small improvements over MFCCs, especially in noisy environments, making them the preferred encoding for many ASR systems [66]. We have also confirmed that mel-generalized cepstra have better ASR performance than mel-cepstra [45], [46].

Using the Mandarin corpus (ca. ten hours of speech data) released for the 2009 Blizzard Challenge, we adapted the above average voice models trained on the Mandarin SPEECON corpus and compared them with SA systems trained on speech data recorded in the office environment and SD systems.

This evaluation was done in the formal listening tests for the 2009 Blizzard Challenge. Mandarin synthetic speech was generated for a set of 697 test sentences, including 647 sentences from a news genre (used to evaluate naturalness and similarity) and 50 semantically unpredictable sentences (used to evaluate intelligibility). The evaluations were conducted over a six week period via the internet, and a total of 334 listeners participated. For further details of these evaluations, see [43].

Fig. 10 shows evaluation results of related systems in the 2009 Blizzard Challenge[9]. From the table, we can see the systems using both noisy data and clean data have slightly better MOS and CCR scores than systems using clean data only. However, contrary to the clean data case in the previous section, the differences are not directly supported by the statistical significance check regardless of the three hundred listeners.

In all the evaluation, natural speech was found to be significantly better than other systems. In the MOS evaluation, the SD system was found to be significantly better than both the SA systems. On the other hand in the similarity and Pinyin with tone error rate (PTER) evaluation, the SD system was found to be significantly better than only the SA systems trained on the office data only. Thus, we can indirectly attribute minor improvements to the additional use of noisy data.

However, in the MOS evaluation, there is a clear gap between the SD system and SA systems, although the amount of noisy data and clean data obtained from the Mandarin SPEECON corpus was 34 hours, which is 3 times more than the amount for the SD-HMMs. This is different from our previous analysis results [5], [28] which show that the SA systems trained on large scale of TTS databases are comparable to the SD systems trained even on eight to ten hours of speech data.

From these results we conclude that the noisy data is not useless. However, mixing speech data recorded in various conditions for ASR is not as efficient as increasing very clean speech data for TTS.

VI. Conclusion

In conventional speaker-dependent speech synthesis including unit-selection and HTS, large amounts of phonetically balanced speech data recorded in highly controlled recording studio environments have typically been required to build a voice. Although using such data is (and will be) a straightforward solution for the best quality synthesis, the number of voices available is always limited, simply because recording costs are high.

On the other hand, in the framework of speaker-adaptive HMM-based speech synthesis, we can consider robust voice building on "non-TTS" corpora such as ASR speech databases. Building TTS voices on ASR speech databases allows the creation of many more voices than has previously been possible for TTS. In fact, we have created the largest collection of

[9]Although the total of 12 Mandarin systems were evaluated in the listening tests, we omitted unrelated systems from this figure and table. For full results, see [43].

synthetic voices in existence from a number of recognized, publicly available ASR corpora—WSJ0, WSJ1, WSJCAM0, RM GlobalPhone, SPEECON, and JNAS. These voices are efficient in terms of total footprints of voices, compared to speaker-dependent HMM systems.

These ASR databases are different from normal purpose-built TTS databases in various ways. Each individual speaker typically records a very limited number of utterances. However, they include hundreds of speakers having various regional accents. Since they are not trained voice talents (who are normally used in TTS databases), session by session their speaking style may undergo sufficient changes to cause issues in TTS systems. The recording condition may not be perfectly consistent and the environment may also vary. Therefore, building TTS voices from these ASR corpora is, in itself, a new challenge.

However, we conclude that relatively clean ASR corpora are very useful, especially for the training of the average voice models used in speaker-adaptive HMM-based speech synthesis because of their rich context coverage. For example the SI-284 system trained on WSJ0 and WSJ1 databases and adapted to TTS databases provided good performance close to that of the reference system trained on ideal TTS databases. Since a lot of clean ASR databases have already been developed for many languages, this result would remove barriers in constructing speaker-adaptive HTS systems for new languages and also would enhance the potential for a unified ASR and TTS framework. The average voice models trained across many speakers themselves have surprisingly high MOS scores. Interestingly, the scores are higher than scores for voices adapted to individual speakers. We believe the average voice models themselves have more potential usefulness than we had initially anticipated and therefore a more complete perceptual and psychological analysis of the average voice models used directly as synthesis models is essential to this approach to speech synthesis.

Contrary to this, the additional use of speech data recorded in various environments such as car or public space resulted in minor improvements. It was not as efficient as increasing the amount of very clean speech data for TTS. The acoustic differences due to the inconsistent recording conditions were also found to be greater than acoustic differences between speakers.

Our evaluation results also show that speaker adaptation on speech data chosen from the ASR corpora presents some difficulty due to the changeable speaking styles, conditions, etc. It would require both a larger amount of speech data than that required for speaker adaptation on TTS corpora and more sophisticated adaptation strategies such as preselection or preclustering of adaptation data.

Meanwhile, from the analysis using many speakers adapted and their average voice available in the ASR corpora, we were able to make new and useful findings. For instance, the MOS scores of the adapted voices were found to be somewhat correlated inversely with mel-cepstral distance from the average voice that the speaker adaptation starts from. Although the correlation is not strong, this becomes an important factor for determining how to train average voice models from many speakers. For instance, this could explain why gender-dependent average voice models provide better speaker adaptation performance than either gender-independent average voice models or speaker-dependent models for TTS. Thus, further in-depth analysis of the relation between the average voice and adapted voices using other acoustic distances such as transformed F_0 and duration distances, or using stochastic measures such as likelihood and Kullback–Leibler divergence is the most important of the future work to have arisen from this study. The analysis results of the distances correlated with the quality of synthetic speech adapted would lead to appropriate speaker clustering for average voice model training.

In the demonstrations of HTS voices built on each ASR corpus, we utilized predefined training sets for each corpus and formed solid benchmarks that have good connections to the ASR field. However, these benchmark systems do not represent the best quality of synthetic speech as our past and new analysis results suggest and some readers would have a strong interest in seeing the highest possible quality being achieved.

For achieving a better quality of synthetic speech based on our analysis results, we should combine these relatively clean ASR corpora and train larger and gender-dependent average voice models. If a huge amount of data is available, we may use multiple gender-dependent average voice models and may choose the nearest model. We note that all of them must have a sufficient quantity of training data since the amount of data for the average voice models is the most dominant factor for the quality of synthetic speech. For the clustering of speakers used for the multiple average voice models, combining the above notions of distances correlated with the quality of synthetic speech adapted and speaker recognition/identification research would be useful. On the other hand, mixing different corpora normally introduces additional differences due to the use of different microphones. This important factor is not well understood and thus requires analysis, which we plan to perform as future work.

We have also shown attractive applications of the voices using a geographical map. In addition to this application, there are several applications which could potentially benefit from the availability of thousands of TTS voices. In closing, we give some practical examples below:

Platform for medical voice banking: These voices may be used as a platform for medical voice banking. In [67], the HTS framework was used as personalized synthetic voices for patients who have dysarthria and thus require TTS systems as communication aids. The patients can choose the most similar voice from a wide variety of voices. Such a selection is most important for patients who start to have progressive speech loss since the amount of speech data available from the patients is very limited and thus adaptation performance highly depends on the initial model from which adaptation will start.

Virtual games and social network services: An individual user can choose a different voice and can avoid overlap of voices between users on virtual games such as second life. For social network services these voices may be attractive.

Forestalling imposture against speaker verification: It is known that the HMM-based speech synthesis system—especially the speaker adaptive framework—can be used to breach speaker verification systems [68], [69]. By using these various kinds of voices, we can verify the robustness of speaker verification systems against imposture using speech synthesis for many speakers in advance [70].

New research field "voice selection for TTS": Finally a new research topic arises from these voices, that is, automatic selection of voices. One such possible solution would be to use GPS as mentioned earlier. Alternatively, from given texts we may estimate an appropriate voice or required conditions (e.g., gender, adult or child, or country, etc.) to be used to read the texts.

ACKNOWLEDGMENT

The authors would like to thank R. B. Chicote of Universidad Politécnica de Madrid and A. Suni of University of Helsinki for the permission of the use of their Spanish and Finnish TTS front-end. They would also like to thank V. Karaiskos for this contribution to our listening tests and the three anonymous reviewers for their constructive feedback and helpful suggestions.

REFERENCES

[1] J. Yamagishi *et al.*, "Thousands of voices for HMM-based speech synthesis," in *Proc. Interspeech-99*, Brighton, U.K., Sep. 2009, pp. 420–423.

[2] 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*, Budapest, Hungary, Sep. 1999, pp. 2374–2350.

[3] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A. B. Black, and T. Nose, "The HMM-Based Speech Synthesis System (HTS) Version 2.1." [Online]. Available: http://hts.sp.nitech.ac.jp/

[4] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. 6th ISCA Workshop Speech Synth. (SSW-6)*, Bonn, Germany, Aug. 2007.

[5] 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, Lang. Process.*, vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

[6] J. Yamagishi, Z.-H. Ling, and S. King, "Robustness of HMM-based speech synthesis," in *Proc. Interspeech–08*, Brisbane, Australia, Sep. 2008, pp. 581–584.

[7] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[8] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. Workshop Speech Natural Lang.*, Harriman, NY, 1992, pp. 357–362.

[9] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," in *Proc. ICASSP-95*, Detroit, MI, May 1995, pp. 81–84.

[10] D. S. Pallet, J. G. Fiscus, and J. S. Garofolo, "DARPA resource management bench," in *Proc. Workshop Speech Natural Lang.*, Hidden Valley, PA, Jun. 1990, pp. 298–305.

[11] T. Schultz, "GlobalPhone: A multilingual speech and text database developed at Karlsruhe university," in *Proc. ICSLP'02*, Denver, CO, Sep. 2002, pp. 345–348.

[12] D. Iskra, B. Grosskopf, K. Marasek, H. V. D. Heuvel, F. Diehl, and A. Kiessling, "SPEECON—speech databases for consumer devices: Database specification and validation," in *Proc. LREC'02*, Canary Islands, Spain, May 2002, pp. 329–333.

[13] M. J. F. Gales and S. J. Young, "The application of hidden Markov models in speech recognition," *Foundations Trends R Signal Process.*, vol. 1, no. 3, pp. 195–304, 2008.

[14] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus," in *Proc. ICSLP-98*, Sydney, Australia, Dec. 1998, pp. 3261–3264.

[15] H. Kubozono, "Mora and syllable," in *The Handbook of Japanese Linguistics*, N. Tsujimura, Ed. New York: Blackwell, 1995, pp. 31–61.

[16] S. Fitt and S. Isard, "Synthesis of regional English using a keyword lexicon," in *Proc. Eurospeech-99*, Budapest, Hungary, Sep. 1999, vol. 2, pp. 823–826.

[17] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Commun.*, vol. 49, no. 4, pp. 317–330, 2007.

[18] [Online]. Available: http://www.lc-star.com

[19] Y. Qian, F. Soong, Y. Chen, and M. Chu, "An HMM-based Mandarin Chinese text-to-speech system," in *Proc. ISCSLP'06*, Singapore, Dec. 2006, pp. 223–232.

[20] Deliverable Report D2.1 EMIME Project, 2008.

[21] Y. Guan, J. Tian, Y.-J. Wu, J. Yamagishi, and J. Nurminen, "An efficient and unified approach of Mandarin HTS system," in *Proc. ICASSP'10*, Dallas, TX, Mar. 2010.

[22] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "XIMERA: A new TTS from ATR based on corpus-based technologies," in *Proc. ISCA 5th Speech Synth. Workshop*, Pittsburgh, PA, Jun. 2004, pp. 179–184.

[23] H. Kawai, T. Toda, J. Yamagishi, T. Hirai, J. Ni, N. Nishizawa, M. Tsuzaki, and K. Tokuda, "XIMERA: A concatenative speech synthesis system with large scale corpora," *IEICE Trans. Inf. Syst.*, vol. J89-D-II, no. 12, pp. 2688–2698, Dec. 2006.

[24] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The ATR multilingual speech-to-speech translation system," *IEEE Trans. Speech, Audio, Lang. Process.*, vol. 14, no. 2, pp. 365–376, Mar. 2006.

[25] R. Barra-Chicote, J. Yamagishi, J. Montero, S. King, S. Lutfi, and J. Macias-Guarasa, "Generacion de una voz sintetica en Castellano basada en HSMM para la Evaluacion Albayzin 2008: Conversion texto a voz," in *V Jornadas en Tecnologia del Habla* (in Spanish), Bilbao, Spain, Nov. 2008, pp. 115–118 [Online]. Available: http://www.cstr.inf.ed.ac.uk/downloads/publications/2008/tts-jth08.pdf

[26] K. Tokuda, H. Zen, and A. W. Black, "HMM-based approach to multilingual speech synthesis," in *Text to Speech Synthesis: New Paradigms and Advances*, S. Narayanan and A. Alwan, Eds. Upper Saddle River, NJ: Prentice-Hall, 2004.

[27] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modeling," in *Proc. ARPA Human Lang. Technol. Workshop*, Plainsboro, NJ, Mar. 1994, pp. 307–312.

[28] J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda, and K. Tokuda, "The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge," in *Proc. Blizzard Challenge 2008*, Brisbane, Australia, Sep. 2008.

[29] M. Fraser and S. King, "The Blizzard Challenge 2007," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Bonn, Germany, Aug. 2007.

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

[31] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, pp. 187–207, 1999.

[32] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.

[33] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 825–834, May 2007.

[34] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc., Series B*, vol. 39, no. 1, pp. 1–38, 1977.

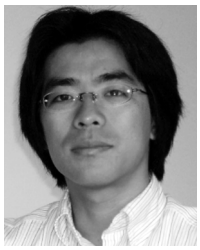
[35] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Japan (E)*, vol. 21, pp. 79–86, Mar. 2000.

[36] Y. Guan and J. Tian, "Evaluation of flat start labeling for phoneme based Mandarin HTS system," in *Proc. ORIENTAL-COCOSDA-09*, Aug. 2009, pp. 187–190.

[37] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP-96*, Philadelphia, PA, Oct. 1996, pp. 1137–1140.

[38] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput. Speech Lang.*, vol. 12, no. 2, pp. 75–98, 1998.

[39] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Speech, Audio, Lang. Process.*, vol. 17, no. 1, pp. 66–83, Jan. 2009.
[40] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.
[41] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5–6, pp. 453–468, 1990.
[42] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP-92*, San Francisco, CA, Mar. 1992, pp. 137–140.
[43] S. King and V. Karaiskos, "The Blizzard Challenge 2009," in *Proc. Blizzard Challenge Workshop*, Edinburgh, U.K., Sep. 2009.
[44] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.
[45] J. Dines, J. Yamagishi, and S. King, "Measuring the gap between HMM-based ASR and TTS," in *Proc. Interspeech-09*, Brighton, U.K., Sep. 2009, pp. 1391–1394.
[46] J. Dines, J. Yamagishi, and S. King, "Measuring the gap between HMM-based ASR and TTS," *IEEE J. Sel. Topics Signal Process.*, 2010, to be published.
[47] D. S. Pallett, J. G. Fiscus, W. M. Fisher, J. S. Garofolo, B. A. Lund, and M. A. Przybocki, "1993 benchmark tests for the ARPA spoken language program," in *Proc. HLT '94: Workshop Human Lang. Technol.*, Morristown, NJ, 1994, pp. 49–74.
[48] Y.-J. Wu, S. King, and K. Tokuda, "Cross-lingual speaker adaptation for HMM-based speech synthesis," in *Proc. ISCSLP-08*, Kunming, China, 2008, pp. 9–12.
[49] Y.-J. Wu and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in *Proc. Interspeech-09*, Brighton, U.K., Sep. 2009, pp. 528–531.
[50] J. A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 5, pp. 380–391, Oct. 1976.
[51] T. P. Barnwell, III, "Correlation analysis of subjective and objective measures for speech quality," in *Proc. ICASSP-80*, Denver, CO, 1980, pp. 706–709.
[52] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features," in *Proc. ICASSP-96*, Atlanta, GA, May 1996, pp. 389–392.
[53] A. W. Black and J. Kominek, "Optimizing segment label boundaries for statistical speech synthesis," in *Proc. ICASSP-09*, Taipei, Taiwan, Apr. 2009, pp. 3785–3788.
[54] R. A. J. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Bonn, Germany, Aug. 2007.
[55] M. Shozakai and G. Nagino, "Analysis of speaking styles by two-dimensional visualization of aggregate of acoustic models," in *Proc. ICSLP-04*, Jeju Island, Korea, Oct. 2004, pp. 717–720.
[56] A. Maier, M. Schuster, U. Eysholdt, T. Haderlein, T. Cincarek, S. Steidl, A. Batliner, S. Wenhardt, and E. Noth, "QMOS—A robust visualization method for speaker dependencies with different microphones," *J. Pattern Recognition Res.*, vol. 1, pp. 32–51, 2009.
[57] T. Cox and M. Cox, *Multidimensional Scaling*. London, U.K.: Chapman & Hall, 2001.
[58] S. Tsakalidis and W. Byrne, "Acoustic training from heterogeneous data sources: Experiments in Mandarin conversational telephone speech transcription," in *Proc. ICASSP-05*, 18–23, 2005, vol. 1, pp. 461–464.
[59] S. Tsakalidis and W. Byrne, "Cross-corpus normalization of diverse acoustic training data for robust HMM training," Cambridge Univ. Eng. Dept., Cambridge, U.K., 2005.
[60] K. Tokuda, H. Zen, and T. Kitamura, "Reformulating the HMM as a trajectory model," *IEICE Tech. Rep. Natural Lang. Understanding Models of Commun.*, vol. 104, no. 538, pp. 43–48, Dec. 2004.
[61] D. Erro, "Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models," Ph.D. dissertation, Univ. Politecnica de Catalunya, Barcelona, Spain, 2008.
[62] J. H. Langlois and L. A. Roggman, "Attractive faces are only average," *Psychol. Sci.*, vol. 1, no. 2, pp. 115–121, 1990.
[63] J. Yamagishi, M. Lincoln, S. King, J. Dines, M. Gibson, J. Tian, and Y. Guan, "Analysis of unsupervised and noise-robust speaker-adaptive HMM-based speech synthesis systems toward a unified ASR and TTS framework," in *Proc. Blizzard Challenge Workshop*, Edinburgh, U.K., Sep. 2009.
[64] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis—A unified approach to speech spectral estimation," in *Proc. ICSLP-94*, Yokohama, Japan, Sep. 1994, pp. 1043–1046.
[65] M. J. F. Gales, "Adaptive training for robust ASR," in *Proc. IEEE Workshop Autom. Speech Recognition Understanding*, Madonna di Campiglio, Italy, 2001, pp. 15–20.
[66] P. C. Woodland, "The development of the HTK broadcast news transcription system: An overview," *Speech Commun.*, vol. 37, no. 1–2, pp. 47–67, 2002.
[67] S. Creer, P. Green, S. Cunningham, and J. Yamagishi, "Building personalised synthesised voices for individuals with dysarthria using the HTS toolkit," in *Computer Synthesized Speech Technologies: Tools for Aiding Impairment*, J. W. Mullennix and S. E. Stern, Eds. Hershey, PA: IGI Global, Jan. 2010.
[68] T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi, "On the security of HMM-based speaker verification systems against imposture using synthetic speech," in *Proc. Eurospeech-99*, Budapest, Hungary, Sep. 1999, pp. 1223–1226.
[69] T. Masuko, K. Tokuda, and T. Kobayashi, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," in *Proc. ICSLP-00*, Beijing, China, Oct. 2000, pp. 302–305.
[70] P. L. De Leon, V. R. Apsingekar, M. Pucher, and J. Yamagishi, "Revisiting the security of speaker verification systems against imposture using synthetic speech," in *Proc. ICASSP-10*, Dallas, TX, Mar. 2010.



Junichi Yamagishi (M'10) received the B.E. degree in computer science and the M.E. and Ph.D. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 2002, 2003, and 2006, respectively.

He held a research fellowship from the Japan Society for the Promotion of Science (JSPS) from 2004 to 2007. He was an Intern Researcher at ATR Spoken Language Communication Research Laboratories (ATR-SLC) from 2003 to 2006. He was a Visiting Researcher at the Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, U.K., from 2006 to 2007. He is currently a Senior Research Fellow at the CSTR, University of Edinburgh, and continues the research on the speaker adaptation for HMM-based speech synthesis in an EC FP7 collaborative project called the *EMIME* project (www.emime.org). His research interests include speech synthesis, speech analysis, and speech recognition.

Dr. Yamagishi pioneered the use of speaker adaptation techniques in HMM-based speech synthesis in his doctoral dissertation *Average-voice-based speech synthesis*, which won the Tejima Doctoral Dissertation Award 2007. He is a member of ISCA, IEICE, and ASJ.

Bela Usabaev received the B.A. degree in computational linguistics from the University of Tübingen, Tübingen, Germany, in 2006. She is currently pursuing the M.S. degree at the University of Tübingen.

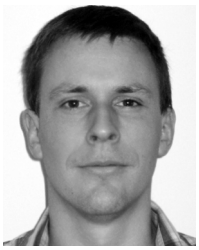
She was an Intern Researcher on HMM-based speech synthesis at Toshiba and CSTR in 2007 and 2008, respectively.

Simon King (M'95–SM'08) received the M.A. (Cantab) and M.Phil. degrees in engineering from the University of Cambridge, Cambridge, U.K., in 1992 and 1993, respectively, and the Ph.D. degree from the University of Edinburgh, Edinburgh, U.K., in 1998.

He is a Reader in linguistics and English language and his interests include speech synthesis, recognition, and signal processing.

Dr. King serves on the ISCA SynSIG committee, co-organizes the Blizzard Challenge, was recently an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, is on the IEEE SLTC, and the editorial board of *Computer Speech and Language*.

Oliver Watts received the M.Sc. degree in Speech and Language Processing from the University of Edinburgh, Edinburgh, U.K., in 2007. He is currently pursuing the Ph.D. degree at the Centre for Speech Technology Research, University of Edinburgh, working on speech synthesis in languages where few resources are available.



John Dines (M'99) graduated with first class honors in electrical and electronic engineering from University of Southern Queensland, Toowoomba, Australia, in 1998 and received the Ph.D. degree from the Queensland University of Technology, Brisbane, Australia, in 2003 with the dissertation "Model based trainable speech synthesis and its applications."

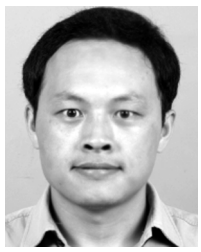
Since 2003, he has been with the Idiap Research Institute, Martigny, Switzerland, where he has been working mostly in the domain of meeting room speech recognition. A major focus of his current research is combining his background in speech recognition and speech synthesis to further advance technologies in both domains.

Dr. Dines is a reviewer for the IEEE SIGNAL PROCESSING LETTERS and the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.

Jilei Tian received the B.S. and M.S. degrees in biomedical engineering from Xi'an Jiaotong University, Xi'an, China, and the Ph.D. degree in computer science from the University of Kuopio, Kuopio, Finland, in 1985, 1988, and 1997, respectively.

He was with the Beijing Jiaotong University faculty from 1988 to 1994. He has been with Nokia Research Center, Beijing, China, as a Senior Research Engineer since 1997, and recently as Principal Member of Research Staff and Team Leader. He has authored or coauthored over 60 refereed publications including book chapter, journal, and conference papers, and holds 40 granted and pending patents. His research interests include speech and natural language processing, human user interface, data mining, context modeling, and biomedical signal processing.

Dr. Tian has served as member of technical committee and technical review committee for conferences and workshops, including ICSLP, Eurospeech, IEEE conferences, etc.

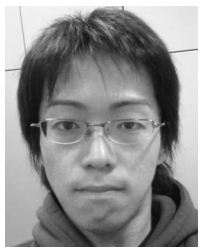


Yong Guan received the B.S. degree from Tsinghua University, Beijing, China, in 2002 and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, in 2008.

He is currently a Postdoctoral Researcher at the Nokia Research Center, Beijing. His research interest covers HMM-based speech synthesis, speech separation, and robust speech/speaker recognition.

Rile Hu received the B.S. and M.S. degrees from the Department of Dynamic Engineering, North China Electric Power University, Beijing, in 1998 and 2001, respectively, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, in 2005.

He is currently a Member of Research Staff at the Nokia Research Center, Beijing. His research interests are in natural language processing, machine learning, and data mining.



Keiichiro Oura received the B.E. degree in computer science and the M.E. and Ph.D. degrees in computer science and engineering from the Nagoya Institute of technology, Nagoya, Japan, in 2005, 2007, and 2010, respectively.

He was an Intern/Co-Op Researcher at ATR Spoken Language Translation Research Laboratories (ATR-SLT), Kyoto, Japan, from September 2007 to December 2007. From April to May 2009, he was a Visiting Researcher in the Centre of Speech Technology Research, Edinburgh, U.K. He is currently a Postdoctoral Fellow of the EMIME project at Nagoya Institute of Technology, Nagoya, Japan. His research interests include statistical speech recognition and synthesis.

Dr. Oura received the Best Student Paper Award at ISCSLP in 2008. He is a student member of the Acoustical Society of Japan.

Yi-Jian Wu received the B.E., M.E., and Ph.D. degrees in electrical engineering and information science from University of Science and Technology of China (USTC), Hefei, Anhui, China, in 2001, 2003, and 2006, respectively.

From April 2003 to March 2004, he was an Intern Student at ATR Spoken Language Translation Research Laboratories (ATRSLT), Kyoto, Japan. He was an Associate Researcher with the Speech Group, Microsoft Research Asia (MSRA) during 2006–2007, and was a Postdoctoral Researcher at the Nagoya Institute of Technology from 2007 to 2009. He is currently an SDE in the TTS Group, Microsoft Business Division, China. His research interests include speech synthesis and speech recognition.

Keiichi Tokuda (M'89) received the B.E. degree in electrical and electronic engineering from the Nagoya Institute of Technology, Nagoya, Japan, and the M.E. and Dr.Eng. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 1984, 1986, and 1989, respectively.

From 1989 to 1996, he was a Research Associate at the Department of Electronic and Electric Engineering, Tokyo Institute of Technology. From 1996 to 2004, he was an Associate Professor at the Department of Computer Science, Nagoya Institute of Technology, as Associate Professor, and currently he is a Professor at the same

institute. He is also an Invited Researcher at ATR Spoken Language Translation Research Laboratories, Japan, and was a Visiting Researcher at Carnegie Mellon University, Pittsburgh, PA, from 2001 to 2002.

Prof. Tokuda is a corecipient of the Paper Award and the Inose Award both from the IEICE in 2001, and the TELECOM System Technology Prize from the Telecommunications Advancement Foundation Award, Japan, in 2001. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society. His research interests include speech coding, speech synthesis and recognition, and statistical machine learning. He is a member of ISCA, IPSJ, IEICE, ASJ, and JSAI.

Reima Karhila is currently pursuing the M.Sc. degree at the Helsinki University of Technology (HUT), Espoo, Finland. His thesis topic and research interests are related to multilingual automatic speech recognition and speaker adaptation.

He is currently a Research Assistant at HUT.

Mikko Kurimo (SM'07) received the Dr.Sc. (Ph.D.) in technology degree in computer science from Helsinki University of Technology, Espoo, Finland, in 1997.

He is currently a Chief Research Scientist and Adjunct Professor at the Department of Information and Computer Science, (previously Laboratory of Computer and Information Science), Helsinki University of Technology. His research interests are in machine learning, speech recognition, information retrieval, natural language processing, and multimodal interfaces.