Analysis of unsupervised and noise-robust speaker-adaptive HMM-based speech synthesis systems toward a unified ASR and TTS framework

Junichi Yamagishi, Mike Lincoln, Simon King, John Dines, Matthew Gibson, Jilei Tian, and Yong Guan. Analysis of unsupervised and noise-robust speaker-adaptive HMM-based speech synthesis systems toward a unified ASR and TTS framework. In Proc. Interspeech 2009, Edinburgh, U.K., September 2009.

Analysis of Unsupervised and Noise-Robust Speaker-Adaptive HMM-Based Speech Synthesis Systems toward a Unified ASR and TTS Framework

Junichi Yamagishi[1], *Mike Lincoln*[1], *Simon King*[1], *John Dines*[2], *Matthew Gibson*[3], *Jilei Tian*[4], *Yong Guan*[4]

[1]University of Edinburgh [2]Idiap Research Institute
[3]University of Cambridge [4]Nokia Research Center

jyamagis@inf.ed.ac.uk

Abstract

For the 2009 Blizzard Challenge we have built an unsupervised version of the HTS-2008 speaker-adaptive HMM-based speech synthesis system for English, and a noise robust version of the systems for Mandarin. They are designed from a multidisciplinary application point of view in that we attempt to integrate the components of the TTS system with other technologies such as ASR. All the average voice models are trained exclusively from recognized, publicly available, ASR databases. Multi-pass LVCSR and confidence scores calculated from confusion network are used for the unsupervised systems, and noisy data recorded in cars or public spaces is used for the noise robust system. We believe the developed systems form solid benchmarks and provide good connections to ASR fields. This paper describes the development of the systems and reports the results and analysis of their evaluation.

Index Terms: speech synthesis, HMMs, speaker adaptation

1. Introduction

Speaker adaptation that transforms a given set of HMMs to a target speaker or condition is a successful technique for both automatic speech recognition (ASR) and HMM-based text-to-speech (TTS) synthesis. [1]. Although we have mainly developed speaker-adaptive HMM-based speech synthesis systems for purely TTS purposes, that is, to improve the similarity of the synthetic speech to natural speech, we can also consider development of these systems from a multidisciplinary application point of view.

Speech-to-Speech Translation (S2ST) that "enables real-time, interpersonal communication via natural spoken language for people who do not share a common language" [2] is a challenging multidisciplinary application for speech processing, and many large-scale projects (Babylon, TC/LC-STAR, EU-Trans, ATR, etc.) have focused on this topic. In our recently-started FP7 project, Effective Multilingual Interaction in Mobile Environments (EMIME) [3], we are developing a device that performs personalized S2ST, such that the user's spoken input in one language is used to produce spoken output in another language, while continuing to sound like the user's voice. Contrary to previous 'pipeline' S2ST systems that combined isolated ASR, machine translation (MT), and TTS systems, or systems that coupled ASR with MT [4, 5], EMIME places major emphasis on coupling ASR with TTS, specifically to simultaneously enable robust, rapid, unsupervised, and cross-lingual speaker adaptation for HMM-based ASR and TTS systems.

The principal modeling framework of speaker-adaptive HMM-based speech synthesis is conceptually similar to conventional ASR systems (but without discriminative training, such as minimum phone error (MPE) [6]) and it is therefore possible to share Gaussians, decision trees or linear transforms between the two [7]. Within this framework we can also consider building TTS voices using ASR corpora [8]. This data is not ideal from a TTS perspective since it may be contaminated with noise, is often recorded under a variety of conditions with microphones of varying quality, and/or may lack phonetic balance. Our recent experiments, however, have demonstrated that speaker-adaptive HMM-based speech synthesis (which uses an 'average voice model' plus speaker adaptation) is robust to the non-ideal data contained in ASR corpora [9]. This naturally leads to a more unified approach that shares noise robust HMMs between ASR and TTS. For noise robust ASR, HMMs are usually trained on speech data corrupted with various kinds of noise, and the systems employ a variety of noise reduction or suppression techniques such as Wiener filtering. For the 2009 challenge we investigate whether it is feasible to train HMMs for TTS from ASR databases that include very noisy speech data (for example recorded in a car or public space) as a first step towards unifying noise robust ASR and TTS. 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.

Another essential task for such an application is to perform the speaker adaptation in an unsupervised manner, allowing completely automatic voice building from arbitrary speech data. In the EMIME project we have developed unsupervised adaptation techniques for HMM-based TTS using either a phoneme recognizer [10], a word-based large-vocabulary continuous speech recognizer (LVCSR) or a technique that maps TTS-HMMs to ASR-HMMs [11]. These are emerging techniques which are undergoing continual improvement, and we therefore investigate the performance of the unsupervised adaptation frameworks to assess what improvements are required for them to compete with supervised TTS systems. For this purpose, we discard all transcriptions and all supervised materials included in the provided corpus and adopt the latest multi-pass large-vocabulary continuous speech recognizer to generate transcriptions of the data.

The EMIME TTS systems are based on the framework from the "HTS-2007 / 2008" system [9, 12], which was a speaker-adaptive system entered for the Blizzard Challenge 2007 and 2008 [13]. We have built an unsupervised version of the HTS-2008 systems for English and a noise robust version of the sys-

tems for Mandarin. Given the more difficult restrictions we impose, our goals are not to improve the quality of the synthetic speech but to make it comparable to that of the original HTS-2008 systems which was trained on supervised clean data. Specifically there are two major aspects that we want to analyze from the Blizzard Challenge, that is, 1) the relationship between the amounts of adaptation data and the performance differences between unsupervised and supervised systems and 2) the usefulness of noisy data. In the following sections we describe the development of the systems and report the results and analysis of their evaluation.

This paper is organized as follows. Section 2 gives an overview and analysis of the ASR corpora used for building EMIME TTS systems. A brief overview of the HTS-2008 speaker-adaptive HMM-based speech synthesis system is given in Section 3.1. Section 3.2 mentions new features such as AMI RT06 LVCSR and adaptation data pruning based on confidence scores used for the unsupervised system. Section 3.3 mentions the use of mel-generalized cepstra for the noise-robust system. System details and performance are described in Section 4. Then section 5 concludes the paper by briefly summarizing our findings.

2. External ASR speech databases used for training of average voice models

In the 2008 challenges, high-quality TTS databases including 40 hours of speech data recorded in highly-controlled recording studio environments were used for training the average voice models in the HTS-2008 system. Of the challenge entrants, the system had equal best naturalness on the small English data set and equal best intelligibility on both small and large data sets [12]. The easiest solution to improve the quality of synthetic speech is to increase the size of the databases, since 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 [1].

For the 2009 challenge, we designed the speaker-adaptive HMM-based speech synthesis systems from a multidisciplinary application point of view, and attempted to integrate the components of the TTS system with other technologies such as ASR. We therefore build the average voice models exclusively from well known and publicly available ASR databases. We believe that the systems form a good benchmark since all the materials used, including databases and tools (e.g. HTK/HTS), are publicly available, and that the techniques provide good connections to ASR fields, encouraging ASR researchers to participate in future Blizzard Challenges.

2.1. English databases

For the English average voice models, we have used the Resource Management database (RM) [14] and the Wall Street Journal databases (WSJ0, WSJ1, and a British English version of WSJ0 called WSJCAM0) [15, 16]. These ASR databases are relatively old, and quite small, typically consisting of 10's of hours of speech, whereas recent ASR systems use thousands of hours of training data. However, they are still typical of clean speech databases that have been (and continue to be) used for over a decade, and it is therefore worthwhile to obtain TTS results from them. Details of the English ASR corpora are shown in Table 1. The amounts of speech data for RM, WSJ0, WSJCAM0, and WSJ1 are 5 hours, 15 hours, 22 hours, and 66 hours respectively. Although the duration for WSJ1 exceeds that of the databases used for HTS-2008, it includes only American English speech data whereas the target speaker of the 2009 Blizzard Challenge is British. Only WSJCAM0 has British speech data.

Table 1: Details of English ASR corpora used for building average voice models.

Corpus (subset)	speakers	sentences/speaker	sentences
RM (ind_total)	160	40.0	6400
WSJ0 (short)	84	86.1	7236
WSJCAM0 (total)	140	81.5	11408
WSJ1 (short)	200	191.3	38278

Table 2: Phonetic coverage of English ASR corpora.

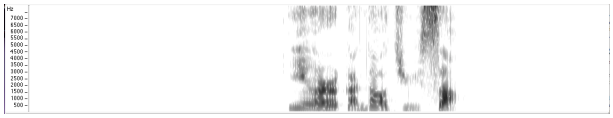
Corpus	triphones/corpus	contexts/corpus
CMU-ARCTIC	10708	91247
RM	7162	114945
WSJ0	18577	421476
WSJCAM0	23534	675266
WSJ0+WSJ1	23776	1246728

Contrary to normal TTS databases where professional or semi-professional narrators utilize standard accents and speaking styles, the speakers included in the ASR databases have a variety of accents. Since the Unilex pronunciation lexicon 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. In practice, however, time constraints meant we were unable to do this, and we simply used general American (GAM) and British received pronunciation (RP) phonesets based on the speaker's nationality. Using speech recordings that comprised a variety of accents for training could be prove to be an advantage or a disadvantage: 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 RP, as it is in this Blizzard Challenge, a more appropriate average voice model would be one trained only on RP speakers, rather than one trained on various accents.

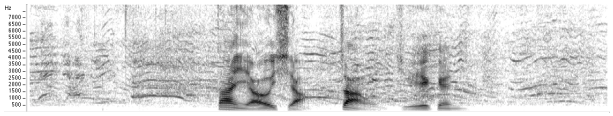
One clear advantage of the ASR corpora is phonetic coverage. Triphone and context coverage is a simple way to measure the phonetic coverage of a corpus. Table 2 shows the total number of different triphone and context types in the English corpora. Since the pre-defined official training data set (known as SI284) for WSJ1 includes WSJ0 as a part of training data, we followed instructions for the November 93 CSR evaluations and calculated them together. 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. 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 sys-

Table 3: Details of the Mandarin Speecon corpus used for building average voice models.

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



(a) Clean data recorded in office space



(b) Noisy data recorded in public space

Figure 1: Spectrograms of clean and noisy data

tems 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.

2.2. Mandarin database

For the Mandarin average voice models, we have used the Mandarin Speecon databases [17]. The Mandarin speecon databases include speech data recorded with various amounts of background noise, detailed below. Sample spectrograms are shown in Figure 1. We directly cite definitions of the noise categories from [17]:

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
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).

Details of each environment are shown in Table 3. 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. Noise levels in dB [A] for each environment are shown in Table 4. We can see that the public space and car environments have larger means and variances. 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 and compared it with systems using all data regardless of the environment. Note that these systems each have about three times as much speech data as the baseline system. For the same reasons as the English system, we used an identical phoneme set for all speakers available in

Table 4: Noise level in dB [A] for each environment.

	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 5: Phonetic coverage of the Mandarin speecon corpus.

Environment	triphones/corpus	contexts/corpus
Office	4999	71863
All environments	5865	181338

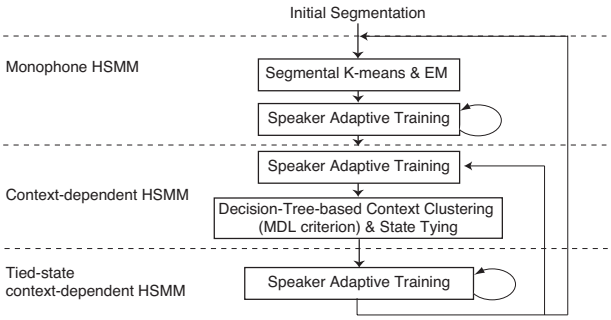


Figure 2: Overview of the training stages of average voice models.

the corpus, although it includes 4 major dialectal accents (Beijing, Chongqing, Shanghai, and Provinces). Table 5 shows the total number of different triphone and context types in the corpus. We see the data set for the mixed environments has a much larger coverage than that of the office environment. There is a trade-off between consistency of recording conditions and phonetic coverage.

The data also includes isolated words, spelling pronunciation utterances and phonetically balanced sentences. Since we are unsure of the effects of using large quantities of isolated words or spelling pronunciation utterances on synthesis, we used only the phonetically balanced sentences as training data for the average voice model in this experiment.

3. EMIME Systems

3.1. HTS-2008: Speaker-adaptive HTS benchmark systems

The HTS-2008 systems utilized speaker-independent HMMs that model acoustic features used for the STRAIGHT [18] mel-cepstral vocoder with mixed excitation (the mel-cepstrum, $\log F_0$ and band-limited aperiodicity measures) [19]. For details see [9, 12].

An overview of the training stages for the average voice models is shown in Figure 2. First, speaker-independent monophone MSD-HSMMs are trained from an initial segmentation, converted into context-dependent MSD-HSMMs, and re-estimated. Then, decision-tree-based context clustering is applied to the HSMMs and the model parameters of the HSMMs are tied. The clustered HSMMs are re-estimated 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. All re-estimation

Table 6: WERs [%] obtained from the AMI 2006 RT system for each genre and pass.

	Genre					
Pass	Address	Arctic	Carroll	Herald 1	Herald 2	Herald 3
P1	47.3	41.7	58.4	40.2	36.6	34.3
P3	41.0	25.5	47.9	26.8	23.5	23.3
P6	40.8	27.5	47.8	28.1	24.8	24.2

and re-segmentation processes utilize speaker-adaptive training (SAT) [20] based on CMLLR [21]. Finally the trained average voice model is adapted to a target speaker in a supervised manner, that is, with the correct labels.

3.2. Unsupervised HTS-2008 using multi-pass LVCSR

3.2.1. AMI RT06 LVCSR

For unsupervised speaker adaptation, we need to automatically obtain transcriptions/labels of the given adaptation data. For this purpose we attempted LVCSR using a system trained on the databases mentioned earlier, and an external LVCSR system called "AMI 2006 RT" system [22].

The first LVCSR system was built by using a technique that maps TTS-HMMs to ASR-HMMs [11] and the 5k/20k language models associated with the WSJ databases. Unfortunately, these language models do not match the ARCTIC sentences selected from Gutenberg which we were required to use for the challenge, and therefore WERs obtained from the LVCSR were poor. Compared to this, the open-domain six-pass AMI 2006 RT system with the 50k language model provided consistently low WERs for various genres including ARCTIC. The six passes of the decoding process are as follows [22]:

P1
Voice activity detection and speaker diarization using GMMs and HMMs [23] followed by initial decoding using a WFST based decoder. [24]

P1 post-processing
Feature level processing such as warping factor estimation for VTLN [25], estimation of LCRC posterior features [26] followed by PCA and HLDA [27], and tandem feature generation from the frequency warped PLP and posterior features.

P2
CMLLR estimation using initial hypothesis obtained from P1, followed by decoding using MPE/VTLN/SAT models with bigram

P3
Lattice expansion (bigram to 4-gram) and rescoring

P4
CMLLR estimation using rescored 1-best hypothesis

P5
MLLR estimation on the top of the CMLLR transforms

P6
Confusion network generation using word posteriors [28, 29]

We chunked the pre-split waveforms for each genre and again performed voice activity detection and speaker diarization. The diarization system uses a unsupervised clustering approach which automatically determines "who spoke when", which includes the estimation of the number of speakers. The diarization of the full Roger database suggested that it included two speakers, however, the second speaker was assigned only a few utterances and thus we excluded the utterances assigned to the second speaker and only utilized those assigned to the first. The WERs of the AMI system for each genre and pass are shown in Table 6 where we see that the WER, even for Arctic, is less than 30% for the final pass. Although the results of the final pass P6 are worse than P3's, we directly utilized 1-best hypothesis of P6 since we wanted to use confidence scores calculated from the confusion network.

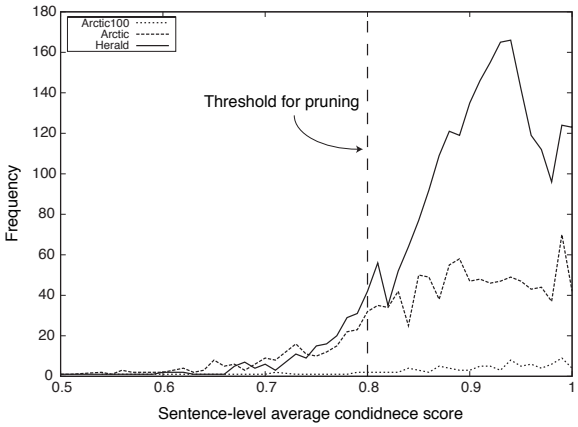


Figure 3: Histogram of weighted confidence scores for each genre.

3.2.2. Adaptation data pruning based on confidence scores

In general the WERs vary by sentence, with some sentences having extremely high WERs. For speaker adaptation it is reasonable to remove data with bad transcriptions. To determine which sentences to prune, we used sentence-level confidence scores obtained from confusion networks in a similar way to [30]. In the original paper, the confidence scores were used to choose poor data that required manual transcription, whereas we use the scores to simply remove poorly transcribed data.

The sentence-level confidences are found by calculating the weighted average of the word level confidence scores as follows:

$$C_s = \frac{\sum_{w \in s} C_w \, T_w}{\sum_{w \in s} T_w} \tag{1}$$

where C_s the sentence-level confidence score of a sentence s, C_w is the word-level confidence score of a word w, and T_w is duration of the word w.

Figure 3 shows histograms of the sentence-level confidence scores for a hundred sentences of the Arctic genre, all sentences of the Arctic genre, and all sentences of the Herald genre. We see that this tails to about 0.5 in terms of confidence scores. The following are examples of automatically obtained transcriptions with high and low confidence scores:

(2) C_s=0.9 Corr: A quarter of Londoners admit the traditional family knees-up always ends in a barney
Auto: A call from London is a bit the traditional family knees up always ends in a bunny

C_s=0.6 Corr: For services to Ophthalmology.
Auto: For services in a form all the G.

Although we have not yet confirmed a correlation between the confidence scores and the quality of synthetic speech, we found that some artifacts can be avoided by pruning data having low confidence scores based on manually specified thresholds for C_s.

3.3. Noise robust HTS-2008

Since this is our first challenge using speech with background car noise etc. for TTS, we did not use any noise suppression techniques. Instead we simply changed acoustic features from mel-cepstra to mel-generalized cepstra [31] with cubic-root compression of amplitude and applied SAT which includes CMN and CVN implicitly and trained the average voice models as normal. Mel-generalized cepstra are similar to PLP features in terms of spectral representation [31]. Thus, we expect that they should provide similar 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 [32]. We have also confirmed that mel-generalized cepstra have better ASR performance than mel-cepstra [7].

3.4. Audio examples

Audio examples for each system above are available online via http://homepages.inf.ed.ac.uk/jyamagis/blizzard09/.

4. System details and their performance

4.1. Released databases / Adaptation data

For the 2009 challenge, a British English and a Mandarin corpora were released. They included 15 hours and 10.5 hours of clean speech data, respectively. Using these corpora and the average voice models above, we built ES1 (100 sentences), EH2 (arctic sentence), and EH1 (all sentences) for English and MS1 (100 sentences) and MH1 (all sentences) for Mandarin.

Supervised training of systems trained on clean data are utilized as speaker-adaptive HTS benchmark (system D) and systems trained either in an unsupervised fashion or on noisy data are utilized as EMIME systems (system W). By comparing these systems, we can see 1) the relationship between the amounts of adaptation data and the performance differences between unsupervised and supervised systems from English results and 2) the usefulness of noisy data from Mandarin results.

4.2. Front-end text processing

The English labels, including the initial segmentation for the data, were automatically generated from the word transcriptions and speech data using the Unisyn lexicon [33] and Festival's Multisyn Build modules [34]. 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 Mandarin labels were also automatically generated from the word transcriptions and speech data using an extended LC-STAR lexica [35] and Nokia's in-house TTS modules based on SAMPA-C. We used phonemes instead of typical Mandarin units, initial/final [36] since we found that the phoneme-based systems outperform when the amount of adaptation data available is limited because of a lower number of units [37].

4.3. Training systems and footprint

We used Edinburgh's grid computing system that has 1456 processors to train average voice models and their adaptations. All the procedures were concurrently conducted per state, per stream, per speaker, and/or per subset of training data. Open MPI further divided them to small subsets using multiple threads.

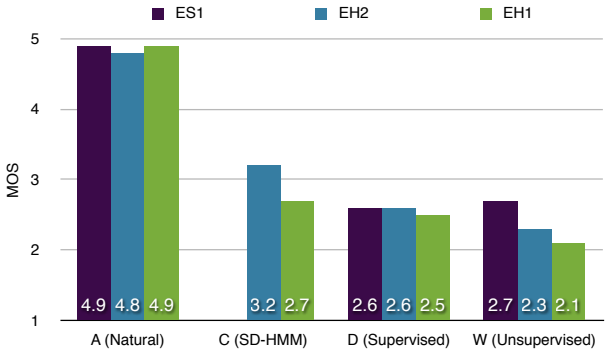


Figure 4: MOS for English EMIME systems

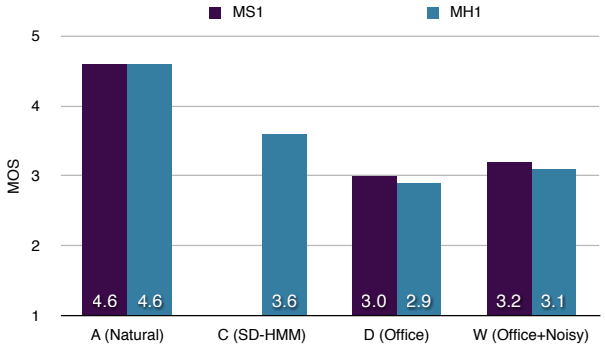


Figure 5: MOS for Mandarin EMIME systems

Tables 7 and 8 show the number of leaves of each of the decision trees and the footprints of each system. From the tables, we see that the numbers of leaves of mel-cepstral and aperiodicity parts for WSJ0 and WSJCAM0 are lower than those of speaker-dependent (SD) HMMs although WSJ0 and WSJCAM0 databases have similar sizes to those for the SD-HMMs. Contrary to this, they have almost the same or more leaves for $\log F_0$ and duration parts. The fact that they include various dialectal accents may partially explain the existence of the redundant $\log F_0$ leaves. However, further investigation is required for the extremely redundant duration leaves. For the Mandarin systems, we see that the system using data in all environments has more leaves than that of the SD-HMMs or that using office environments only.

4.4. Findings from the 2009 Blizzard Challenge results

Synthetic speech was generated for a set of 445 test sentences, including sentences from conversational, news and novel genres (used to evaluate naturalness and similarity) and semantically unpredictable sentences (used to evaluate intelligibility). To evaluate naturalness and similarity, 5-point mean opinion score (MOS) and CCR tests were conducted. The scale for the MOS test was 5 for "completely natural" and 1 for "completely unnatural". The scale for the CCR tests was 5 for "sounds like exactly the same person" and 1 for "sounds like a totally different person" compared to natural example sentences from the reference speaker (EM001). To evaluate intelligibility, the subjects were asked to transcribe semantically unpredictable sentences; average word error rates (WER) were calculated from these transcripts. The evaluations were conducted over a six

Table 7: The number of leaf nodes and footprints for each English speaker-dependent (SD) and speaker-adaptive system.

	Number of leaves in decision tree				Footprint of acoustic models [MB]	
System	Mel-cepstrum	$\log F_0$	Aperiodicity	Duration	HTK format	hts_engine format
SD (full)	5833	27137	6790	4045	517	13.0
RM	2122	12417	2839	3733	334	5.8
WSJ0	2945	26952	2624	13165	669	11.0
WSJCAM0	3599	40326	3237	23641	981	13.0
WSJ0+WSJ1	10861	105940	9202	51567	1697	34.0

Table 8: The number of leaf nodes and footprints for each Mandarin speaker-dependent (SD) and speaker-adaptive system.

	Number of leaves in decision tree				Footprint of acoustic models [MB]	
System	Mel-cepstrum	$\log F_0$	Aperiodicity	Duration	HTK format	hts_engine format
SD (full)	4837	14175	4654	3335	400	9.5
Office environment	2373	15320	3238	3474	442	7.0
All environments	6272	33905	6378	7695	681	17.0

week period via the internet, and a total of 482 and 334 listeners participated for English and Mandarin, respectively. For further details of these evaluations, see [38].

Figures 4 and 5 show MOS results of related systems in the Blizzard evaluation. We summarize our findings below:

The unsupervised approach (D vs W in English)
The unsupervised systems are comparable to the supervised system when the amount of adaptation data is limited. This is consistent with our previous experiments [11, 39]. However they become less comparable as the amount of available data increases.

The use of noisy data (D vs W in Mandarin)
The systems using both noisy data and clean data have comparable or slightly better MOS values than systems using clean data only. This is a promising result since mixing noisy data would be essential for noise robust ASR and this implies that TTS can share the HMMs.

We can see the same tendency from results for both similarity and intelligibility tests.

4.5. Differences between 2008 and 2009 results

By comparing evaluation results of the speaker-dependent and speaker-adaptive HTS systems in the 2008 and 2009 Blizzard Challenge (C vs V in 2008 challenge [12] and C vs D in 2009 challenge), we can see the effect of different average voice models indirectly.

In the 2008 challenge, the HTS-2008 systems using 40 hours of TTS speech data outperforms SD-HMMs. However the 2009 system trained on ASR corpora do not reach the quality of SD-HMMs in either of the two languages. The fact that HMMs trained on the WSJCAM0 database is as small as ones trained on the target speaker's database (see Table 7) may mostly explain the performance reduction in English. On the other hand, the amount of noisy data and clean data obtained from the Mandarin ASR corpora was 34 hours, which is 3 times more than the amount for the SD-HMMs. However there is a clear gap between systems C and W in the 2009 Mandarin results. From this we conclude that using speech data recorded in various conditions for ASR is not as efficient as increasing very clean speech data for TTS.

5. Follow-up study

Since we did not have enough time to analyze the proposed systems during the Blizzard Challenge period, we performed follow-up listening tests for evaluating the unsupervised approach including the use of confidence score mentioned in Sect. 3.2 and our hypothesis mentioned in Sect. 4.5.

System configurations used for the follow-up listening tests are shown in Table 9. The first group (System A to E) was designed for comparison of speech databases used for training of the average voice models. "CSTR" represents CSTR in-house 40 hours of TTS speech data used for the HTS-2008 systems in the 2008 challenge. The second group (D, F to J) was designed for comparison of the unsupervised approach and the use of confidence score. "P1" and "P6" refer to the transcriptions automatically obtained in the P1 and P6 recognition pass of the AMI RT06 LVCSR system. "CS Threshold" represents the threshold for pruning adaptation data based on the confidence scores. We utilized all test sentences of the 2007, 2008, and 2009 challenge and evaluate naturalness and similarity in a similar way to one of the Blizzard Challenge. The number of listeners who completed the listening test was 68.

The evaluation results are also shown in the same table. In the MOS evaluation on naturalness for the first group, system A was found to be statistically better than other systems except system E ($p < 0.01$). The system E was also found to statistically better than system G ($p < 0.01$). Other differences in the group were not statistically significant. In the similarity (SIM) evaluation for the same group, there was no statistically significant difference.

In summary the ASR corpus that should have been used for training of the average voice models in this Blizzard Challenge was not the British English WSJCAM0 database (system D) but the American English WSJ0+WSJ1 database (system E) even for the British target speaker, because the database has about 80 hours of speech data. This underpins our hypothesis on the performance reduction mentioned in Sect. 4.5. This result also underpins our another hypothesis on the effect of various recording conditions of ASR corpora, since the system trained on the WSJ0+WSJ1 database does not outperform the original HTS-2008 system (system A) that uses 40 hours of TTS speech data, which is half amount of the WSJ0+WSJ1 database.

Table 9: System configurations used for follow-up listening tests. CS stands for confidence score. MOS and similarity scores are also given in the last two columns.

Index	Corpus	Duration (h)	Supervision	Transcriptions	CS Threshold	MOS	SIM
A	CSTR	40	Y	Multisyn	n/a	3.4	3.1
B	RM	5	Y	Multisyn	n/a	2.2	2.6
C	WSJ0	15	Y	Multisyn	n/a	2.6	2.9
D	WSJCAM0	22	Y	Multisyn	n/a	2.7	2.7
E	WSJ0+WSJ1	81	Y	Multisyn	n/a	2.8	3.0
D	WSJCAM0	22	Y	Multisyn	n/a	2.8	3.0
F	WSJCAM0	22	N	P1	0.00	2.3	2.5
G	WSJCAM0	22	N	P6	0.00	2.5	2.5
H	WSJCAM0	22	N	P6	0.80	2.4	2.6
I	WSJCAM0	22	N	P6	0.90	2.4	2.5
J	WSJCAM0	22	N	P6	0.95	2.4	2.4

In both the MOS and similarity evaluation for the second group, there was no statistically significant difference observed. Since the difference between supervised and unsupervised systems are relatively small, we need to collect more listeners.

6. Conclusions

For the 2009 Blizzard Challenge we have built an unsupervised version of the HTS-2008 speaker-adaptive HMM-based speech synthesis system for English and a noise robust version of the system for Mandarin and have analyzed their performance.

The unsupervised approach and noisy data may be used in the framework of speaker-adaptive HMM-based speech synthesis system to strongly link ASR and TTS. However, they also have some negative effects for TTS: the unsupervised approach are comparable to the supervised system when the amount of adaptation data is limited, but they become less comparable as the amount of available data increases. The systems using both noisy data and clean data have comparable or slightly better MOS values than systems using clean data only, but the use of speech data recorded in various conditions for ASR is not as efficient as increasing very clean speech data for TTS.

7. Acknowledgements

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

8. References

[1] 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 & Language Process.*, vol. 17, no. 1, pp. 66–83, 1 2009.

[2] F. Liu, L. Gu, Y. Gao, and M. Picheny, "Use of statistical N-gram models in natural language generation for machine translation," in *Proc. ICASSP 2003*, Hong Kong, Apr. 2003, pp. 636–639.

[3] [Online]. Available: http://www.emime.org

[4] H. Ney, "Speech translation: coupling of recognition and translation," in *Proc. ICASSP-99*, Phoenix, Arizona, Mar. 1999, pp. 517–520.

[5] Y. Gao, "Coupling vs. unifying: Modeling techniques for speech-to-speech translation," in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, Sept. 2003, pp. 365–368.

[6] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP 2002*, Orlando, Florida, May 2002, pp. 13–17.

[7] J. Dines, J. Yamagishi, and S. King, "Measuring the gap between HMM-based ASR and TTS," in *Proc. Interspeech 2009*, Brighton, U.K., Sept. 2009, (to be appear).

[8] J. Yamagishi *et al.*, "Thousands of voices for HMM-based speech synthesis," in *Proc. Interspeech 2009*, Brighton, U.K., Sept. 2009, (to be appear).

[9] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "A robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. Speech, Audio & Language Process.*, vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

[10] S. King, K. Tokuda, H. Zen, and J. Yamagishi, "Unsupervised adaptation for HMM-based speech synthesis," in *Proc. Interspeech 2008*, Brisbane, Australia, Sept. 2008, pp. 1869–1872.

[11] M. Gibson, "Two-pass decision tree construction for unsupervised adaptation of HMM-based synthesis models," in *Proc. Interspeech 2009*, Brighton, U.K., Sept. 2009, (to be appear).

[12] 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, Sept. 2008.

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

[14] D. S. Pallet, J. G. Fiscus, and J. S. Garofolo, "DARPA resource management benchmark test results June 1990," in *Proceedings of the workshop on Speech and Natural Language*, Hidden Valley, Pennsylvania, 1990, pp. 298–305.

[15] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*, Harriman, New York, 1992, pp. 357–362.

[16] 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.

[17] D. Iskra, B. Grosskopf, K. Marasek, H. Heuvel, F. Diehl, and A. Kiessling, "Speecon - speech databases for consumer devices: Database specification and validation," in *Proc. LREC, 2002*, Las Palmas, Spain, 2002, pp. 329–333.

[18] 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 Communication*, vol. 27, pp. 187–207, 1999.

[19] 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.

[20] 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.

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

[22] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiát, D. van Leeuwen, M. Lincoln, and V. Wan, "The 2007 AMI(DA) system for meeting transcription," in *CLEAR*, 2007, pp. 414–428.

[23] D. van Leeuwen and M. Huijbregts, "The AMI speaker diarization system for NIST RT06s meeting data," in *Machine Learning for Multimodal Interaction*, ser. Lecture Notes in Computer Science, vol. 4299. Berlin: Springer Verlag, October 2007, pp. 371–384.

[24] D. Moore, J. Dines, M. Magimai.-Doss, J. Vepa, O. Cheng, and T. Hain, "Juicer: A weighted finite-state transducer speech decoder," in *3rd Joint Workshop on Multimodal Interaction and Related Machine LEarning Algorithms MLMI'06*, 2006, IDIAP-RR 06-21.

[25] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. ICASSP 1996*, Atlanta, Georgia, May 1999, pp. 346–348.

[26] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognitio," in *Proc. ICASSP 2006*, Toulouse, France, May 2006, pp. 325–328.

[27] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 26, no. 4, pp. 283 – 297, 1998.

[28] G. Evermann, P. Woodland, and P. C. Woodl, "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *Proc. ICASSP 2000*, 2000, pp. 2366–2369.

[29] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 373 – 400, 2000.

[30] K. Yu, M. J. F. Gales, and P. C. Woodland, "Unsupervised training with directed manual transcription for recognising mandarin broadcast audio," in *Proc. Interspeech 2007*, Antwerp, Belgium, 2007, pp. 1709–1712.

[31] 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, Sept. 1994, pp. 1043–1046.

[32] P. C. Woodland, "The development of the HTK broadcast news transcription system: An overview," *Speech Communication*, vol. 37, no. 1-2, pp. 47 – 67, 2002.

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

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

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

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

[37] EMIME Project, "Deliverable report D2.1," November 2008, (Available on request).

[38] S. King and V. Karaiskos, "The Blizzard Challenge 2009," in *Proc. Blizzard Challenge Workshop*, Edinburgh, U.K., Sept. 2009.

[39] J. Dines, J. Yamagishi, and S. King, "Measuring the gap between HMM-based ASR and TTS," *IEEE Journal of Selected Topics in Signal Processing*, vol. Submitted, September 2009.