IDIAP RESEARCH REPORT

MEASURING THE GAP BETWEEN HMM-BASED ASR AND TTS

John Dines Junichi Yamagishi Simon King

Idiap-RR-34-2010

OCTOBER 2010

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Measuring the Gap Between HMM-Based ASR and TTS

John Dines, *Member, IEEE*, Junichi Yamagishi, *Member, IEEE*, Simon King, *Senior Member, IEEE*

Abstract

The EMIME European project is conducting research in the development of technologies for mobile, personalised speech-to-speech translation systems. The hidden Markov model (HMM) is being used as the underlying technology in both automatic speech recognition (ASR) and text-to-speech synthesis (TTS) components, thus, the investigation of unified statistical modelling approaches has become an implicit goal of our research. As one of the first steps towards this goal, we have been investigating commonalities and differences between HMM-based ASR and TTS. In this paper we present results and analysis of a series of experiments that have been conducted on English ASR and TTS systems measuring their performance with respect to phone set and lexicon; acoustic feature type and dimensionality; HMM topology; and speaker adaptation. Our results show that, although the fundamental statistical model may be essentially the same, optimal ASR and TTS performance often demands diametrically opposed system designs. This represents a major challenge to be addressed in the investigation of such unified modelling approaches.

Index Terms: speech synthesis, speech recognition, unified models

I. Introduction

Over the last decade automatic speech recognition (ASR) and text-to-speech synthesis (TTS) technologies have shown a convergence towards statistical parametric approaches [1]–[3] . Despite this apparent convergence of technologies, the ASR and TTS communities continue to conduct their research in a largely independent fashion, with occasional cross-overs between the two. On one hand this can be considered a natural consequence of the fact that these technologies have quite disparate goals in mind, but it can also be argued that there are several persuasive arguments for considering ASR and TTS technologies in a more unified context.

A core motivation for conducting research in the domain of unified speech modelling is the possibility of better understanding the mathematical and theoretical relationship between synthesis and recognition. Furthermore, this

John Dines* is with the IDIAP Research Institute, Centre du Parc Martigny Switzerland. Tel. +41-27-721-77-11 Fax +41-27-721-77-12 E-mail: john.dines@idiap.ch. *Corresponding author.

Junichi Yamagishi and Simon King are with the Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, EH8 9AB, United Kingdom. TEL: +44-131-650-4434 FAX: +44-131-650-6626 E-mail: jyamagis@inf.ed.ac.uk, simon.king@ed.ac.uk.

Manuscript received August 11, 2009, revised November 16 2009. 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). SK holds an EPSRC Advanced Research Fellowship. JY is partially supported by EPSRC. This work has made use of the resources provided by the Edinburgh Compute and Data Facility which is partially supported by the eDIKT initiative (http://www.edikt.org.uk). Simplified descriptions of this research are introduced in a paper that appears in the proceedings of Interspeech 2009.

may encourage greater cross-polination of knowledge between the two fields, leading to novel discoveries in both. The last and possibly greatest motivation comes from the possibilities that unified modelling of ASR and TTS offer in terms of applications. Arguably, the application most likely to benefit from unified models for ASR and TTS is that of speech-to-speech translation (SST), which combines ASR, TTS and machine translation (MT).

While several speech-to-speech translation efforts have been conducted over the years, most have used largely heterogeneous approaches[1]. In EMIME[2], we aim to use statistical parametric methods in order to achieve two goals in SST; firstly, the ability to efficiently adapt a system to the user's voice and, secondly, in the context of a mobile application, we wish to benefit from parsimonious nature of such approaches. More specifically, we are using hidden Markov model (HMM) based automatic speech recognition (ASR) and text-to-speech (TTS) in order to achieve these goals.

The use of unified models in SST represents a particularly attractive paradigm since it provides a natural mechanism for speaker-adaptive TTS by employing the same speaker dependent transforms learned from ASR, while offering further efficiency with respect to computation and memory (see for eg. [4]–[6]). There are numerous challenges present in developing such models. In particular we note that, despite the common underlying statistical framework, HMM-based ASR and TTS systems are generally very different in their implementation.

This paper presents a detailed empirical study of ASR and TTS systems, where evaluations are carried out using a common training data set and (where possible) common model training paradigm. Our goal is to determine which components of TTS and ASR systems are the most detrimental to the other, thus, identifying priorities for further research in the development of unified models. Thus, if our ultimate goal is to 'bridge the gap' between ASR and TTS then this work is primarily concerned with 'measuring the gap' between ASR and TTS.

The paper is organised as follows: Section II presents an overview statistical models for ASR and TTS, focusing on the HMM and the major differences between ASR and TTS approaches. Section III describes our methodology and Section IV details our empirical studies and analysis in measuring the the gap between ASR and TTS systems. Finally in Section V we present our conclusions.

II. Statistical generative modelling of speech for ASR and TTS

Automatic speech recognition and text-to-speech synthesis have fundamentally different objectives: ASR is concerned with classification/discrimination of time series and TTS is concerned with generation/regression of time series. In ASR, both generative and discriminative modelling approaches have been extensively investigated. More recently, increasing attention has been made towards the study of discriminative models such as conditional random field (CRF) [7] and discriminative training criteria such as maximum mutual information (MMI) [8] and minimum phone error (MPE) [9], since in classification tasks there is little point in accurately representing the entire observation space when our interest is primarily on the decision boundaries between classes. In contrast for TTS,

[1]For example: Technology and Corpora for Speech to Speech Translation (TC-STAR) http://www.tc-star.org/; Global Autonomous Language Exploitation (GALE) http://www.darpa.mil/ipto/programs/gale/gale.asp; The Vermobil Project http://verbmobil.dfki.de/overview-us.html.

[2]Effective Multilingual Interaction in Mobile Environments: http://www.emime.org

investigations have naturally been limited to generative modelling, though alternative training/generation criteria are also emerging [10].

In considering the different time series statistical models that have been proposed for ASR and TTS, we focus on the generative models. The most extensively investigated generative model has been the hidden Markov model that was first proposed for use in ASR [11] and subsequently for TTS [12]. The HMM only provides a coarse approximation of the underlying process for the generation of acoustic observations, in particular, the conditional independence assumption of acoustic features and the first order Markovian assumption for state transitions. Consequently, numerous models have been proposed that attempt to overcome the short-falls of the HMM and provide better performance with respect to ASR and/or TTS.

The most elementary effort to improve modelling of the HMM has been the inclusion of dynamic features [13], which does not even require modification of the model, but has significant impact on ASR and TTS. Similarly, the hidden semi-Markov model (HSMM) provides explicit modelling of state duration through a simple modification to the HMM that is particularly important for synthesis [14]. Due to the importance of feature dynamics in speech synthesis, the explicit relationship between between dynamic and static features has been exploited during inference of observation vectors [3]. For consistency, this explicit relationship should also be taken into account during model parameter estimation, leading to the development of the trajectory HMM [15], which has been shown to further benefit both ASR and TTS performance.

Aside from the trajectory HMM, alternative generative models have been studied that explicitly model feature dynamics; for example, in the form of: 1) state trend functions [16], [17]; 2) as an auto-regressive process [18], [19]; or 3) at the segment level using switching dynamic system [20], [21]. Implementation of such statistical modelling frameworks for ASR and TTS also requires consideration of the sparse nature of contextual modelling, where some models, such as switching linear dynamic system, are able to provide implicit handling of co-articulation effects resulting in a more parsimonious model, while others constitute a more direct extension of the conventional HMM framework and necessitate a reformulation of parameter tying algorithms [22].

Deep architectures provide a means for efficiently learning complex tasks such as those encountered in speech and language processing. In particular, it is argued that shallow architectures can require exponentially more computational elements than an appropriately deep architecture [23]. Such shallow architectures are typified in conventional ASR and TTS systems that explicitly model conditional distributions of all the contexts. One such deep architecture that is based on a generative framework is the deep belief network (DBN) [24], which has been shown to yield impressive performance on a phone recognition task [25]. An alternative that provides a less dramatic break from conventional modelling approaches includes methods for generating ensembles of trees that can provide a more efficient means to tie acoustic contexts in HMM-based systems [26], [27].

The duality of generative models for both classification and regression tasks provides a basis for unified modelling approaches and motivates us to evaluate such models not only in terms of classification performance for ASR, but also in terms of generation performance; that is, using measures such as spectral distortion and subjective evaluation. Such an in-depth comparison from these different perspectives has the potential to provide more insight into the

Configuration	ASR	TTS
General		
Lexicon	CMU	Unisyn
Phone set	CMU (39 phones)	GAM (56 phones)
Acoustic parameterization		
Spectral analysis	fixed size window	STRAIGHT (F_0 adaptive window)
Feature extraction	filter-bank cepstrum (Δ + Δ^2)	mel-generalised cepstrum (+ Δ + Δ^2) + $\log F_0$ + bndap (+ Δ + Δ^2)
Feature dimensionality	39	120 + 3 + 15
Frame shift	10ms	5ms
Acoustic modelling		
Number of states per model	3	5
Number of streams	1	5
Duration modelling	transition matrix	explicit duration distribution (HSMM)
Parameter tying	phonetic decision tree (HTK)	shared decision tree (MDL)
State emission distribution	16 component GMM	single Gaussian pdf
Context	triphone	full (quinphone + prosody)
Training	2-pass system (ML-SI & ML-SAT)	Average voice (ML-SAT)
Speaker adaptation	CMLLR	CMLLR or CSMAPLR

TABLE I
CONFIGURATIONS OF HMM-BASED ASR AND TTS SYSTEMS.

performance of the generative models. In this paper we limit the scope of our investigations to the dominant paradigm in speech modelling for ASR and TTS – the hidden Markov model. We expect that many of the findings would generalise to other generative models that have been mentioned above.

A. HMM-based ASR and TTS

The hidden Markov model has been the dominant paradigm for ASR for over two decades. In more recent years the HMM has also become the focus of increasing interest in TTS research. This apparent convergence of ASR and TTS to a common statistical parametric modelling framework is largely thanks to a number of properties of the HMM, among these the most notable include its scalability to large scale tasks; desirable generalisation properties; powerful adaptation framework; and parsimony with respect to the size of training data. The continued dominance

of HMM-based techniques is also thanks, in part, to the existence of freely available software such as HTK [28], a trend that is also continuing in TTS with HTS [29]. In comparing typical HMM-based ASR and TTS systems, there are a few fundamental differences that we can note, in particular, unlike in speech recognition, speech synthesis utilises explicit state duration modelling; modelling of semi-continuous data; and makes extensive use of a full range of contextual information for the prediction of prosodic patterns [30], [31].

Less evident, but equally important, are the specifics of how these systems are implemented. Components such as lexicon and phone set, acoustic features, and HMM topology are generally different in ASR and TTS systems, our choice being influenced by the differing goals of ASR and TTS. In the case of ASR, robustness to speaker and environmental variability, ability to handle pronunciation variation and generalisation to unseen data while maximising class discrimination are paramount. In TTS we are concerned with such characteristics as the ability to re-synthesise speech which is highly intelligible and retains speaker identity and also the ability to generate natural sounding speech from previously unseen text. Many of these desirable properties are diametrically opposed, thus we expect many properties of ASR and TTS systems to be incompatible. Table I shows typical configurations of HMM-based ASR and TTS systems (these also being the baseline configurations we have used for experiments described in this paper). For further details of such systems refer to [28], [29], [32].

In the study presented in this paper we analyse ASR and TTS performance with respect to several key system components namely: lexicon and phone set; feature extraction; model topology; and speaker adaptation. This study has been conducted with American-English systems using phone based acoustic units, though we believe that many of the results are also significant for other languages, even when the phoneme is not typically the acoustic unit of choice (see for eg. [33]). In the remainder of this section we present brief descriptions of these components and refer to previous related studies that have been conducted. We note that although some previous experiments have been conducted which compare the aforementioned aspects of ASR and TTS, we believe this is the most comprehensive such study and the first to consider both ASR and TTS.

1) Lexicon and phone set: The lexicon describes the set of words known by the system and their pronunciation(s). In TTS we may also generate pronunciations that lie outside of the lexicon using letter-to-sound (LTS) methods. In practice, lexica can differ greatly, both in terms of the phone set and the way in which phones are composed into word pronunciations. There is no strict set of guidelines as to what constitutes an optimal lexicon for application in either ASR or TTS, though it is evident that in both cases phone sequences produced by the lexicon should have good correlation with acoustic data. There has been significant work conducted on pronunciation variation modelling for ASR [34], but there are few detailed studies investigating the choice of lexicon and phone set for ASR or TTS. One of the few such studies [35] shows that the choice of lexicon can lead to significantly different performance between ASR systems.

2) Feature extraction: Typically, there are significant differences between feature extraction techniques used in ASR and TTS. In recognition, emphasis is placed on speech representations that provide good discrimination between speech sounds, while being relatively invariant to speaker identity and environmental factors. The ability to reconstruct speech from such representations is not necessary, so much information may be discarded. Conversely,

parametric models for synthesis are focused on reconstruction and manipulation of the speech signal, incorporating higher order analysis and a method for signal reconstruction. ASR systems typically employ a filterbank based cepstrum representation such as perceptual linear prediction (PLP) [36]. TTS features are normally based on variations of the mel-generalised cepstrum analysis [37] and may incorporate STRAIGHT F_0-adaptive spectral analysis [38] .

The literature shows numerous studies comparing different feature extraction techniques for ASR and TTS, amongst which we can find work that is particularly relevant to the study reported here [39], [40], though there are few such comparisons that take both ASR and TTS into consideration [41]. Furthermore, studies in ASR have largely been concerned with low order feature analysis, while TTS studies have tended to focus on higher analysis orders. In summarising the findings of this work, we see that in general higher order features are better suited to TTS and lower order features ASR. Unfortunately, there is little information comparing ASR and TTS features on a common task, and the evaluation tasks that have been used are often insufficiently complex or use too little data in order to elucidate significant differences between systems.

3) Model topology: Model topology describes the manner in which states in the HMM set are arranged. Thus, we can consider the number of emitting states in each model as one aspect of model topology as well as the state transition modelling (eg. left-right, ergodic, explicit duration pdf). In ASR, it is typical to employ 3-state left-right HMM topology, whereas in TTS 5-state left-right HSMM topology is normally employed.

We may also consider parameter smoothing and parameter tying techniques, such as decision tree state tying, as being concerned with model topology. Both ASR and TTS use variants of decision tree state tying [42]. Recognition systems are usually built using a single tree per state per base phone (phonetic decision tree), whereas synthesis models tend to use a single tree per state (shared decision tree). Stopping criteria for tree growth are normally either based on minimum likelihood increase combined with a minimum lead node occupancy threshold (as is used in HTK) or use a model selection criteria such as minimum description length (MDL) [43].

Overall, there appears to be a dearth of information in the literature concerning optimal selection of HMM topology, though there has been some work reported on alternatives to the standard left-right configuration [44] and also work showing the link between parameter tying and pronunciation modelling [34]. Within both the ASR and TTS research communities a common HMM topology seems to have been almost unanimously adopted, which suggests that these have been accepted to be the optimal configurations. Concerning state-tying, we can point to previous work [45], [46], which have shown that the MDL criterion works well for clustering without the need to fine tune the system.

4) Speaker adaptation: Arguably, the most pervasive speaker adaptation approaches in speech recognition and speech synthesis are those based on maximum likelihood linear transforms (MLLT) [47] and maximum a posteriori (MAP) adaptation [48] – where the two may also be used in combination [49]. Such approaches provide the means to adjust models using relatively few parameters, thus requiring only a small quantity of speaker-specific data. Several flavours of linear transform-based speaker adaptation exist that may be applied to model parameters (maximum likelihood linear regression (MLLR) [50], structural maximum a posteriori linear regression (SMAPLR)

[51]) or features (constrained maximum likelihood linear regression (CMLLR) [47], constrained structural maximum a posteriori linear regression (CSMAPLR) [46], [52]). Speaker adaptive training (SAT) [53] uses speaker dependent transforms during training of the speaker independent HMM acoustic model, such that the speaker acoustic model is comprised of both the canonical acoustic model plus speaker dependent transforms. SAT has been used extensively in ASR and TTS (where the canonical model is called the *average voice model* [40]).

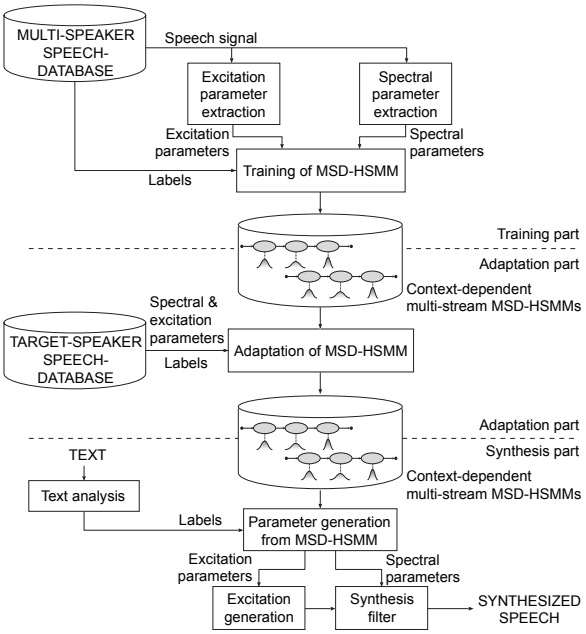
Adaptation may be performed in supervised mode – where we know the transcription of the adaptation data – and in unsupervised mode – where we do not know the true transcription of the adaptation data and adaptation is performed using an estimated ASR transcription. Numerous comparisons of speaker adaptation algorithms have been made for both ASR and TTS comparing adaptation algorithms [46], [47] and supervised versus unsupervised adaptation [4], [50].

III. Methodology

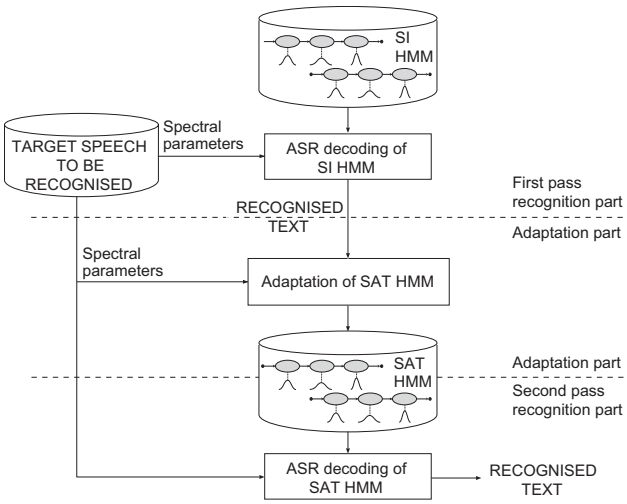
The experiments presented in this paper have been conducted using existing techniques in ASR and TTS. Conventional evaluation measures have been adopted in order to allow comparison with other systems reported in the literature. As far as possible, the variables for experimentation in the TTS evaluations (e.g., training and test sets, speech features, and so on) are shared between both ASR and TTS systems. Since our goal is to understand which aspects of ASR and TTS systems are compatible and those which diverge, the methodology that we have undertaken is to compare ASR and TTS performance for baseline systems against systems where we have exchanged baseline components for those in the opposing system (eg. we exchange ASR features for TTS features and evaluate these in the context of ASR WER and visa versa). The baseline system configurations are shown in Table I. In this study we are not considering such fundamental differences as duration or context modelling – these being the subject of more focused research [27], [54], [55].

A. Experimental setup

We built the ASR and TTS systems based on the HTS system entry to the 2007/2008 Blizzard Challenge [32], [56]. The HTS-2007 system is illustrated in Figure 1a, where four main components can be identified: speech analysis, average voice training, speaker adaptation and speech synthesis. An additional recognition part is illustrated in Figure 1b. The speech analysis stage is responsible for the generation of acoustic features upon which our models are trained. For speech synthesis, speech analysis composes F_0 adaptive STRAIGHT spectral analysis [38] followed by extraction of mel-generalised cepstrum-derived spectral parameters [37] plus excitation parameters ($\log F_0$ and band-limited aperiodic features (bndap) for mixed excitation). Each feature is modelled using a separate stream, where semi-continuous features ($\log F_0 + \Delta + \Delta^2$) use multi-space probability distribution (MSD) [30]. For speech recognition, speech analysis uses perceptual linear prediction (PLP) coefficients [36]. We model only spectral features for the speech recognition component of this study, hence, speech recognition models use a single stream whereas speech synthesis models use five separate feature streams.



(a) Overview of the HTS 2007 speech synthesis system. The system comprises speech analysis, average voice training, speaker adaptation and speech synthesis stages.



(b) Recognition part of the system.

Fig. 1. Overview of ASR and TTS system configuration used in this work.

Speech recognition and synthesis systems use the same average voice training procedure which involves the generation of maximum likelihood speaker adaptive trained (SAT) [53], context dependent, left-right models. The synthesis synthesis system uses only a single diagonal mixture component per state emission pdf. The speech recognition system has its state emission pdfs incremented to 16 diagonal Gaussian mixture components. Duration modelling for the ASR system uses standard transition matrix whereas TTS system uses explicit modelling of state duration using a single Gaussian per state [57]. ASR models use triphone based context with phonetic decision trees,

TTS models use "full-context" (incorporating both quinphone and prosodic context labels) with shared decision trees.

Constrained maximum likelihood linear regression (CMLLR) [47] is used during training and testing of both synthesis and recognition systems. By default, the ASR system uses unsupervised adaptation in a two-pass configuration, using speaker independent models for the first pass and SAT trained models in the second pass. The baseline TTS system uses supervised adaptation. The application of unsupervised adaptation to TTS is a subject of ongoing research [4], [5], which we also touch upon in this study. Synthesis uses HMM-based parameter generation [58], [59] to generate sequences of excitation and spectrum parameters. Excitation parameters are used to generate a source signal using pitch synchronous overlap and add (PSOLA). The speech waveform is generated by exciting a mel-logarithmic approximation filter (MSLA) that corresponds to the generated spectral parameters with the source signal.

Training data comprised the Wall Street Journal (WSJ0) short term speaker training data (SI84) which includes 7240 recordings made by 84 speakers [60]. The use of an ASR corpus for training synthesis models is a new concept though does not involve any technical novelty. Our motivation in doing so was to ensure a maximum of commonality between ASR and TTS systems and thus greater consistency in the reporting of experimental results. Furthermore, our ultimate goal is the development of unified modelling approaches which implies that we use common training data for ASR and TTS. In a separate study we have shown that using ASR corpora to build TTS corpora is indeed a reasonable thing to do [61], [62].

B. ASR evaluation

For the evaluation of ASR we used the primary condition (P0) of the 5k vocabulary hub task (H2) of the November 93 CSR evaluations, except for speaker adaptation evaluations for which we use the Spoke 4 (S4) task of the November 93 CSR evaluations. Decoding employs the 5k closed bigram language model distributed with the corpus. The word error rate (WER) metric is used in the reporting of ASR system performance. Statistical significant testing of ASR results is carried out using the bootstrap method [63] and is reported with 95% confidence.

C. TTS evaluation

For the evaluation of TTS we also used the November 1993 CSR Spoke 4 data. The large number of design factors that can be varied during the training of an HMM-based synthesiser leads to a potentially very large number of variants to be compared. Therefore, listening tests have only been used for a subset of systems, and for a single target speaker, '4oa'. Objective measures have been used for all systems and all the target speakers from the evaluation set. It is important to recognise 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 [64], [65].

Objective evaluation is carried out by first aligning reference and test utterances. To measure the accuracy of the spectral envelope of the synthetic speech, we use "average mel-cepstral distance" (MCD) [66]–[70], which is only

calculated during periods of speech activity. The MCD calculated between the mel-cepstra generated from HMMs and extracted from the natural reference speech in the evaluation set is an Euclidean distance and is given by

$$\text{MCD [dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}(c_d - \hat{c}_d)^2}, \tag{1}$$

where D is the analysis order of mel-cepstra and c_d and $\hat{c}_d$ are the d-th coefficients of the mel-cepstra of generated and natural speech, respectively. Note that the c_0 term which captures the power of waveforms is excluded from the MCD calculation. To measure the accuracy of the F_0 contour, 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 only calculated when both generated and the actual speech are voiced. Lastly, we measure voicing error as the percentage of frames in which the natural and synthetic speech differ in their voicing status.

For subjective evaluation of synthesised speech, we adopted a design based on that of the 2007/2008 Blizzard Challenges [32], [71], which are open evaluations of corpus-based TTS synthesis systems. To evaluate speech naturalness 5-point mean opinion score (MOS) are used. The scale for the MOS test runs from 5 for "completely natural" to 1 for "completely unnatural". 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. The evaluations were conducted via a standard web browser with a total of 124 paid native English speakers participating in these tests.

IV. Results and analysis

This section details experiments conducted for ASR and TTS systems for different system components as described in Section II. For completeness we list all results and corresponding statistical significance in Appendix A. Readers should refer to the appendix for precise details concerning system configurations.

A. Comparison of phone set and lexicon

The CMU lexicon [72] was used in the baseline ASR system and the Unisyn lexicon [73] with general American accent (GAM) in the baseline TTS system. These lexica use phone sets consisting of 39 phones and 56 phone respectively. A version of the Unisyn lexicon using an Arpabet-like set consisting of 45 phonemes was also evaluated. Table II lists the phone sets used in these studies and mappings between the three. The CMU phone set mapping is only approximate, since a one-to-one mapping does not exist due to inconsistencies between the underlying pronunciations in the CMU and Unisyn lexica.

The results of lexicon evaluations are shown in Table III. We can see that the extended GAM phone set leads to a decrease in ASR performance, which can be alleviated through the Arpabet mapping, finally giving superior performance to that of the baseline system. Closer analysis of the GAM phone set shows that a number of the phones may be considered allophones or composites of other phones. These phones have relatively few occurrences in the training data, which may lead to acoustic models of these phones being poorly trained. We note, however,

GAM		Arpabet	CMU	GAM		Arpabet	CMU	GAM		Arpabet	CMU
@	46k	ax	ah, ih	ii	21k	iy	iy	r	42k	r	r
a	10k	ae	ae	ir*	4.5k	iy	ih	@r (+r)	12k	axr	er
aa	9k	aa	aa	jh	3k	jh	jh	@@r (+r)	3k	er	er
aer*	230	ay	ay	k	22k	k	k	s	29k	s	s
ai	7k	ay	ay	l	6k	l	l	sh	5k	sh	s
ar*	2k	aa	aa	l!	4k	el	ah+l	t	34k	t	t
b	10k	b	b	lw*	11k	l	l	tˆ	8k	dx	t,d
ch	2k	ch	ch	m	17k	m	m	th	3k	th	th
d	23k	d	d	m!	70	em	ah+m	u	1k	uh	uh
dh	11k	dh	dh	n	34k	n	n	uh	10k	ah	ah
e	14k	eh	eh	n!	5.5k	en	ah+n	ur*	500	uh	uh
eh*	500	ae	ae	ng	5k	ng	ng	uu	6.5k	uw	uw
ei	9k	ey	ey	oi	1k	oy	oy	v	9k	v	v
eir	1.5k	ey	eh	oo	3k	ao	ao	w	8k	w	w
f	10k	f	f	or	3.5k	ao	ao	y	4k	y	y
g	3.5k	g	g	ou	7.5k	ow	ow	z	17k	z	z
h	6k	hh	hh	ow	2k	aw	aw	zh	309	zh	zh
hw	850	w	w	owr*	270	aw	aw				
i	29k	ih	ih,iy	p	16k	p	p				

TABLE II

PHONE SETS FOR DIFFERENT LEXICA AND THEIR COUNTS ON THE WSJ SI-84 TRAINING DATA (FOR GAM ONLY). GAM PHONES MARKED WITH * ARE MERGED WITH OTHER PHONES IN THE ARPABET PHONE SET.

Lexicon	Phone set (size)	ASR WER (%)	TTS		
			MCD	RMSE of $\log F_0$	V/UV error
CMU	CMU (39)	6.4	5.63	198	16.9
Unisyn	GAM (56)	6.6	<u>5.56</u>	198	<u>15.7</u>
Unisyn	Arpabet (45)	<u>6.1</u>	5.60	198	16.3

TABLE III

COMPARISONS OF LEXICA FOR ASR AND TTS. COMPLETE SYSTEM CONFIGURATIONS CAN BE FOUND IN TABLES 3 AND 4.

that none of the above ASR results were found to be statistically significant. We would need to evaluate with a larger test set in order to confirm the above hypotheses.

Observations for TTS are to the contrary of ASR with the Unisyn lexicon giving slightly better objective measures

Feature		ASR WER			TTS	
Type	Order	All	Male	Female	WER	MOS
PLP	13	6.8	8.2	5.4	–	–
	25	8.1	9.3	6.7	–	–
	40	11.9	11.7	12.2	–	–
MCEP	13	9.4	10.4	8.4	–	–
	25	10.9	12.4	9.3	–	–
	40	19.1	20.0	18.1	–	–
STRAIGHT+MCEP	13	11.4	13.6	9.1	15	1.9
	25	12.8	14.1	11.5	20	2.4
	40	16.0	18.9	12.9	21	2.7
STRAIGHT+MGCEP	13	10.3	12.7	7.9	19	2.0
	25	10.2	12.0	8.3	24	2.5
	40	13.6	15.9	11.2	24	2.3
STRAIGHT+MGLSP	13	–	–	–	18	2.0
	25	–	–	–	16	2.7
	40	–	–	–	19	2.5

TABLE IV

COMPARISONS OF FEATURE CONFIGURATIONS FOR ASR AND TTS. COMPARISONS ARE MADE WITH RESPECT TO FEATURE ANALYSIS ORDER AND FEATURE EXTRACTION METHOD. COMPLETE SYSTEM CONFIGURATIONS CAN BE FOUND IN TABLES 3 AND 4.

in the sense of mel-cepstral distance and V/UV error. We hypothesise that this is derived from the richer labelling of the Unisyn lexicon providing better prediction of allophonic variations. Overall, all systems give very similar results.

B. Comparison of feature extraction

The ASR system uses perceptual linear prediction coefficients (PLP) as the baseline features whereas the TTS system uses features based on mel-generalised cepstral analysis (MGCEP) of STRAIGHT spectrum[3]. More specifically, mel-generalised analysis may be used to derive a cepstral representation using generalised logarithm in which the hyper-parameter, $\gamma = 0$, corresponds to logarithmic compression of the spectrum (STRAIGHT+MCEP) and $\gamma = -1/3$ corresponds to cubed-root spectral compression (STRAIGHT+MGCEP). STRAIGHT+MGLSP analysis corresponds to frequency warped line-spectrum pair parameterisation, in which $\gamma = -1$. Systems have all been trained using the MDL criterion for state tying, obviating the need to explicitly choose a threshold for controlling tree growth. As previously stated, we do not consider features for $\log F_0$ or aperiodicity measures in ASR experiments. The results of these comparisons are shown in Table IV.

[3]Feature normalisation (eg. CMN/CVN) is not used in ASR or TTS systems, this being implicit to feature space adaptation.

First of all, we see that conventional ASR features perform substantially better than any of the TTS mel-ceptrum-based features of equivalent order in the ASR task. One of the main differences between typical ASR features and the MGCEP analysis is the use of filter-banks during frequency warping, hence, we postulate this as a possible reason for their increased robustness since the sum-log operation of the filter-bank can help to reduce sensitivity to frequency bins with low SNR. STRAIGHT spectrum also appears to be detrimental to ASR performance, most likely due to sensitivity to F_0 extraction inaccuracies. Of all of the MGCEP-based features, the STRAIGHT+MGCEP features provide the best performance on average for ASR, which is consistent with results reported in the literature. We also note that MGCEP-based features are closest in terms of signal processing to the PLP features. For TTS, subjective evaluations reveal that there is little to separate the different feature analysis methods.

Concerning feature analysis order, we see that ASR and TTS systems behave in a contrary fashion. ASR performance degrades rapidly as analysis order increases, while TTS quality degrades as order decreases. TTS intelligibility is not significantly affected by analysis order. When considering the most likely explanations for this behaviour it is important to remember that lower order cepstra are generally accepted to contain the most important information for speech sound discrimination, whereas higher order ceptra contain finer details of the spectrum, including information pertaining to speaker identity. This is supported by the fact that speaker identification systems generally also use higher order cepstra [74]. The practical consequence is that ASR systems have their performance degraded when modelling higher order cepstra, as the bulk of information contained therein is irrelevant to the task at hand, and likewise in TTS, the exclusion of higher order cepstra removes much of the information necessary for high quality synthesis and maintaining speaker identity (though not speech intelligibility). Results of particular interest were obtained with the STRAIGHT+MGCEP features at an analysis order of 25, which show the lowest degradation to performance of both ASR and TTS when compared, respectively, to lower and higher analysis orders.

An additional point worth noting from these results concerns the impact of STRAIGHT analysis on ASR performance. We note that STRAIGHT analysis appears to degrade ASR performance at lower analysis orders, but at higher orders it is actually beneficial to ASR. This is due to the ability of the STRAIGHT analysis to remove harmonic components from the spectrum that would otherwise be captured by higher order cepstra. In particular, we observe that the STRAIGHT analysis technique provides improved performance for female speakers due to the greater spacing between harmonics of female speakers[4].

C. Comparison of model topology

We conducted experiments with respect to HMM topology by comparing different state tying schemes, where the ASR baseline uses phonetic decision tree (one tree per phone per state) combined with likelihood and minimum occupancy thresholds to control tree growth whereas the TTS baseline uses shared decision tree (one tree per

[4]This is contrary to what was reported earlier in [75]. Based on the observations of this work we re-conducted the experiments incorporating a more robust F_0 extraction algorithm, more specifically, voiced/unvoiced detection accuracy was greatly improved in order to account for significant gaps of waveform power between training and evaluation data. Voicing detection is important for STRAIGHT analysis, since it is F_0 adaptive and thus V/UV error causes huge differences in its spectral analysis.

Tree type	Clustering	Threshold		ASR	TTS				
		TB	RO	WER	MCD	RMSE of $\log F_0$	V/UV error	MOS	WER
Phonetic	HTK	450	200	9.4	–	–	–	–	–
	MDL	–	–	9.4	5.66	447	15.9	1.5	26
Shared	HTK	300	200	9.4	–	–	–	–	–
	MDL	–	–	9.2	5.56	198	15.7	2.7	21

TABLE V

COMPARISONS OF STATE-TYING FOR ASR AND TTS. THRESHOLDS FOR HTK TREE TYING, TB AND RO, CORRESPOND TO MINIMUM LIKELIHOOD INCREASE AND NODE OCCUPANCY, RESPECTIVELY. ASR SYSTEMS HAVE BEEN TUNED FOR OPTIMAL PERFORMANCE WITH RESPECT TO DECISION TREE GROWTH. COMPLETE SYSTEM CONFIGURATIONS CAN BE FOUND IN TABLES 3 AND 4.

state) with MDL criterion to control tree growth. The phonetic versus shared tree offer their own advantages and disadvantages, in particular, the phonetic decision tree should minimise confusion between phones whereas the shared tree is able to provide more efficient sharing of parameters across models. Table V shows the results of these experiments.

An unexpected result for the ASR experiments revealed that the shared decision tree yielded equivalent performance to that of the phonetic decision tree. Recalling the results for the comparison between lexica, we found that the reduced Arpabet phone set produced lower WER than the original Unisyn phone set. We hypothesise that the shared decision tree is able to perform a similar mapping by clustering models across phone classes that would otherwise remain distinct in the phonetic decision tree, achieving a data-driven reduction of the phone set. However, working against any such benefit gained from sharing across phone classes is the possibility of increased confusability between triphone models with different centre phones. To what extent these two factors affect system performance must depend on the training data, phone sets and lexicon.

The TTS results show that the phonetic decision tree-based tying results in worse performance than shared decision trees, in particular, for the $\log F_0$ feature streams. The HMM used for TTS does not need to discriminate each phoneme perfectly and, particularly for $\log F_0$, sharing models across phone classes allows more effective modelling of supra-segmental effects. In practice, phoneme-based clustering makes little sense for $\log F_0$; in the $\log F_0$ shared trees, stress or accentual categories appear near the root, rather than phone classes.

In order to further analyse the relationship between state clustering approaches and model complexity we conducted a series of ASR experiments in which we tuned the respective thresholds controlling decision tree growth. These experiments were conducted with both MDL and ML stopping criteria; the results are shown in Figure 2. The ASR experiments confirm results previously reported for TTS, where it has been shown that MDL acts as an appropriate criterion for stopping tree growth without the need for time-consuming tuning of hyper-parameters.

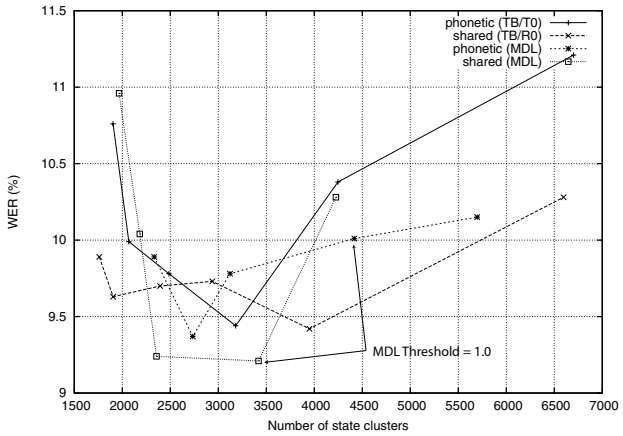


Fig. 2. Analysis of decision tree tuning for ASR. System configuration is the same as that reported in Table V.

Adaptation algorithm	Supervised adaptation	Transcription recognizer	ASR WER	TTS				
				MCD	RMSE of $\log F_0$	V/UV error	MOS	WER
MLLR	Y	–	11.5	5.46	192	11.8	3.0	13
CMLLR	Y	–	13.2	5.50	204	11.8	2.6	19
SMAPLR	Y	–	11.5	5.43	192	11.8	2.9	14
CSMAPLR	Y	–	13.0	5.50	215	11.8	2.6	19
SMAPLR+MAP	Y	–	11.5	5.46	191	16.1	2.8	19
CSMAPLR+MAP	Y	–	13.9	5.56	198	15.7	2.7	21
CSMAPLR+MAP	N	SI 5k-bg	14.3	5.60	203	25.7	2.7	21
CSMAPLR+MAP	N	SI 20k-bg	14.4	5.59	200	25.5	2.6	21
CSMAPLR+MAP	N	SAT 5k-bg	14.3	5.59	201	25.4	2.4	23
CSMAPLR+MAP	N	SAT 20k-bg	14.2	5.58	198	25.5	2.5	23

TABLE VI

EVALUATION OF SPEAKER ADAPTATION. COMPLETE SYSTEM CONFIGURATIONS CAN BE FOUND IN TABLES 3 AND 4.

D. Comparison of Speaker adaptation

We compared speaker adaptation for ASR and TTS with respect to adaptation algorithms and supervised and unsupervised adaptation. For supervised adaptation of ASR, we generated triphone context labels directly from the word-level transcription of the adaptation data. Similarly, for supervised adaptation of TTS we generate full context labels by processing word-level transcriptions using TTS front-end. Adaptation is then performed using the model-level transcriptions. For the evaluation of unsupervised ASR and TTS systems we generate adaptation transforms from the output of ASR systems with various WER, thus enabling assessment of adaptation performance with respect to the degree of noise in the ASR transcription. Unsupervised TTS requires that full-context transcriptions

are generated from word-level ASR output as in the supervised case.

The evaluation of adaptation was carried out using Spoke 4 (S4) task of the November 1993 CSR evaluations. All adaptation was carried out off-line using the rapid enrolment data (for condition 'C3') which comprises 40 adaptation utterances for each of the 4 speakers. For the subjective evaluation of TTS, a single male target speaker was chosen at random from the S4 task, and the 40 block adaptation utterances provided for this speaker were used to adapt the average voice models. As this enrolment data does not lie within the domain of the provided word lists and language models, WERs of the ASR systems on the enrolment data are higher than that which is usually reported for the S4 task itself. WERs were as follows on the enrolment data using speaker independent (SI) and speaker adaptive models (SAT) with 5k and 20k wordlists and bigram language models: SI 5k-bg 59.7%, SAT 5k-bg 41.2%, SI 20k-bg 23.5%, SAT 20k-bg 17.3%. We also measured phone error rate (PER) for these systems of 20.2%, 15.1%, 10.4% and 6.5% respectively. The results of these experiments are shown in Table VI.

We make note of several observations concerning these results. Firstly, it is apparent that for both ASR and TTS there is not statistically significant difference between adaptation algorithms, though the results a slight preference for mean-transform based adaptation over feature transform based adaptation for this task. Furthermore, MAP adaptation does not appear to provide any additional benefits. Secondly, comparing supervised and unsupervised adaptation reveals a small degradation to ASR performance when using unsupervised adaptation, while TTS shows no significant degradation, irrespective of the WER/PER of the underlying transcription. This is a significant result, since it shows that TTS systems can be adapted to a specific person's voice without knowledge of what has been spoken. It is worth pointing out that even when the correct word transcription is available, we cannot be sure the full context labels exactly correspond to the speech signal. This means that even the supervised adaptation is operating with noisy full context labels. This may be part of the reason why the unsupervised systems are no worse than the supervised system (or visa-versa).

V. Conclusions

We have presented a series of 'measuring the gap' experiments exploring the differences between HMM-based ASR and TTS systems. These experiments provide valuable insight to several key challenges towards the development of unified models for ASR and TTS. Our findings in these experiments show that, many of the techniques used in ASR and TTS can not be simply applied to their respective other without negative consequences. In particular, we note the following major findings concerning each of the areas investigated and possible future research directions:

- **Lexicon and phone set:** There is weak evidence suggesting smaller phone sets are favoured by ASR whereas larger phone sets with allophonic variants may be favoured for TTS, but in general no significant differences were found between the different lexica and phone sets that were tested.
- **Feature extraction:** Feature extraction methods used in TTS were found to result in significantly poorer ASR performance than conventional ASR feature extraction. For TTS, no significant differences were measured between different feature extraction methods. Furthermore, higher dimensionality features, as are usually

necessary for high quality waveform generation, were found to significantly degrade ASR, whereas the converse was observed for TTS performance. This result stems from the fact that ASR and TTS rely on different aspects of the spectrum for optimal performance. Of all features compared, STRAIGHT+MGCEP seems to give the best compromise in terms of ASR and TTS performance, with the STRAIGHT analysis being critical to obtaining good performance with high analysis order. Future research needs to concentrate on developing more robust (in ASR terms) spectral analysis methods that still permit high quality signal reconstruction (for TTS), which may include the development of alternative vocoding approaches. Likewise, methods for dimensionality reduction may provide means to improve ASR performance while minimising impact on TTS.

- **Model topology:** Experiments evaluated HMM topology, in particular, parameter tying schemes. ASR results showed that the choice of stopping criterion is not critical given that the system is properly tuned, though the MDL criterion may simplify this process. Surprisingly, ASR results also demonstrated that shared decision tree tying could provide equivalent performance than the phonetic decision tree tying. TTS experiments showed that shared versus phonetic decision tree tying has little impact on spectrum or voicing decision (V/UV), but is critical for prediction of F_0 and duration since they rely on supra-segmental rather than phonetic contexts. Overall, a judicious choice of system configuration should avoid any negative impact on either ASR or TTS performance.
- **Speaker adaptation:** ASR and TTS experiments compared several speaker adaptation algorithms for which it was found that model space transforms were preferred over feature space transforms, though, there was little to separate all algorithms compared. No significant differences were found between the unsupervised and supervised TTS systems in terms of naturalness, similarity or intelligibility. For ASR systems, a small but significant difference was measured between supervised and unsupervised adaptation. Future work in adaptation may follow several directions. Firstly, we noted that limitations of full-context label generation for TTS systems may be a limiting factor with respect to the comparison of unsupervised and supervised adaptation, hence, alternative methods for full-context label generation should be studied. Additionally, both ASR and (even more so) TTS systems are limited by the quantity of adaptation data available to them. Means to rapidly adapt these systems using as little data as a single utterance would also appear to be an interesting research direction.

Additional research topics that may naturally follow on from this work include the investigation of how TTS modelling may contribute to ASR, in terms of the use of full-context models and modelling of excitation features. Furthermore, the investigation of unsupervised adaptation techniques for TTS is a new idea that stands to gain much from closer integration of ASR and TTS methodologies. We expect to see new applications in the near future that leverage from our results, including automatic personalisation of TTS systems, especially in the domain of speech-to-speech translation.

APPENDIX A
COMPLETE LISTING OF RESULTS

We list the full set of ASR and TTS systems evaluated in Tables VII, VIII, IX and X; and Figures 3, 4 and 5.

Index	Training data	Spectral analysis Method	Order	Tree structure	Adaptation algorithm	Supervised adaptation	Transcription recognizer	MOS	WER (%)
A	**SI84**	MCEP	40	shared	CSMAPLR+MAP	Y	–	2.7	21
B	**40H**	MCEP	40	shared	CSMAPLR+MAP	Y	–	2.0	35
C	SI84	**MCEP**	**13**	shared	CSMAPLR+MAP	Y	–	1.9	15
D	SI84	MCEP	**25**	shared	CSMAPLR+MAP	Y	–	2.4	20
A	SI84	MCEP	**40**	shared	CSMAPLR+MAP	Y	–	2.7	21
E	SI84	**MGCEP**	**13**	shared	CSMAPLR+MAP	Y	–	2.0	19
F	SI84	MGCEP	**25**	shared	CSMAPLR+MAP	Y	–	2.5	24
G	SI84	MGCEP	**40**	shared	CSMAPLR+MAP	Y	–	2.3	24
H	SI84	**MGC-LSP**	**13**	shared	CSMAPLR+MAP	Y	–	2.0	18
I	SI84	MGC-LSP	**25**	shared	CSMAPLR+MAP	Y	–	2.7	16
J	SI84	MGC-LSP	**40**	shared	CSMAPLR+MAP	Y	–	2.5	19
A	SI84	MCEP	40	**shared**	CSMAPLR+MAP	Y	–	2.7	21
K	SI84	MCEP	40	**phonetic**	CSMAPLR+MAP	Y	–	1.5	26
L	SI84	MCEP	40	shared	**MLLR**	Y	–	3.0	13
M	SI84	MCEP	40	shared	**CMLLR**	Y	–	2.6	19
N	SI84	MCEP	40	shared	**SMAPLR**	Y	–	2.9	14
O	SI84	MCEP	40	shared	**CSMAPLR**	Y	–	2.6	19
P	SI84	MCEP	40	shared	**SMAPLR+MAP**	Y	–	2.8	19
A	SI84	MCEP	40	shared	**CSMAPLR+MAP**	Y	–	2.7	21
A	SI84	MCEP	40	shared	CSMAPLR+MAP	**Y**	–	2.7	21
Q	SI84	MCEP	40	shared	CSMAPLR+MAP	**N**	**SI 5k-bg**	2.7	21
R	SI84	MCEP	40	shared	CSMAPLR+MAP	**N**	**SI 20k-bg**	2.6	21
S	SI84	MCEP	40	shared	CSMAPLR+MAP	**N**	**SAT 5k-bg**	2.4	23
T	SI84	MCEP	40	shared	CSMAPLR+MAP	**N**	**SAT 20k-bg**	2.5	23

TABLE VII

THE 20 TTS SYSTEMS THAT WERE EVALUATED IN THE LISTENING TEST. SOME ROWS ARE DUPLICATED TO MAKE BETWEEN-SYSTEM COMPARISONS EASIER TO READ. BOLD FACE IS USED TO HIGHLIGHT THE SETTING(S) BEING VARIED IN EACH SUBSET OF RESULTS. TRAINING DATA SET '40H' IS THE 40 HOURS OF DATA USED IN THE HTS ENTRY TO BLIZZARD 2008 [56]. MOS MEANS 'MEDIAN NATURALNESS' AND WER IS INTELLIGIBILITY MEASURED USING SEMANTICALLY UNPREDICTABLE SENTENCES. ALL SYSTEMS USE STRAIGHT SPECTRAL ANALYSIS.

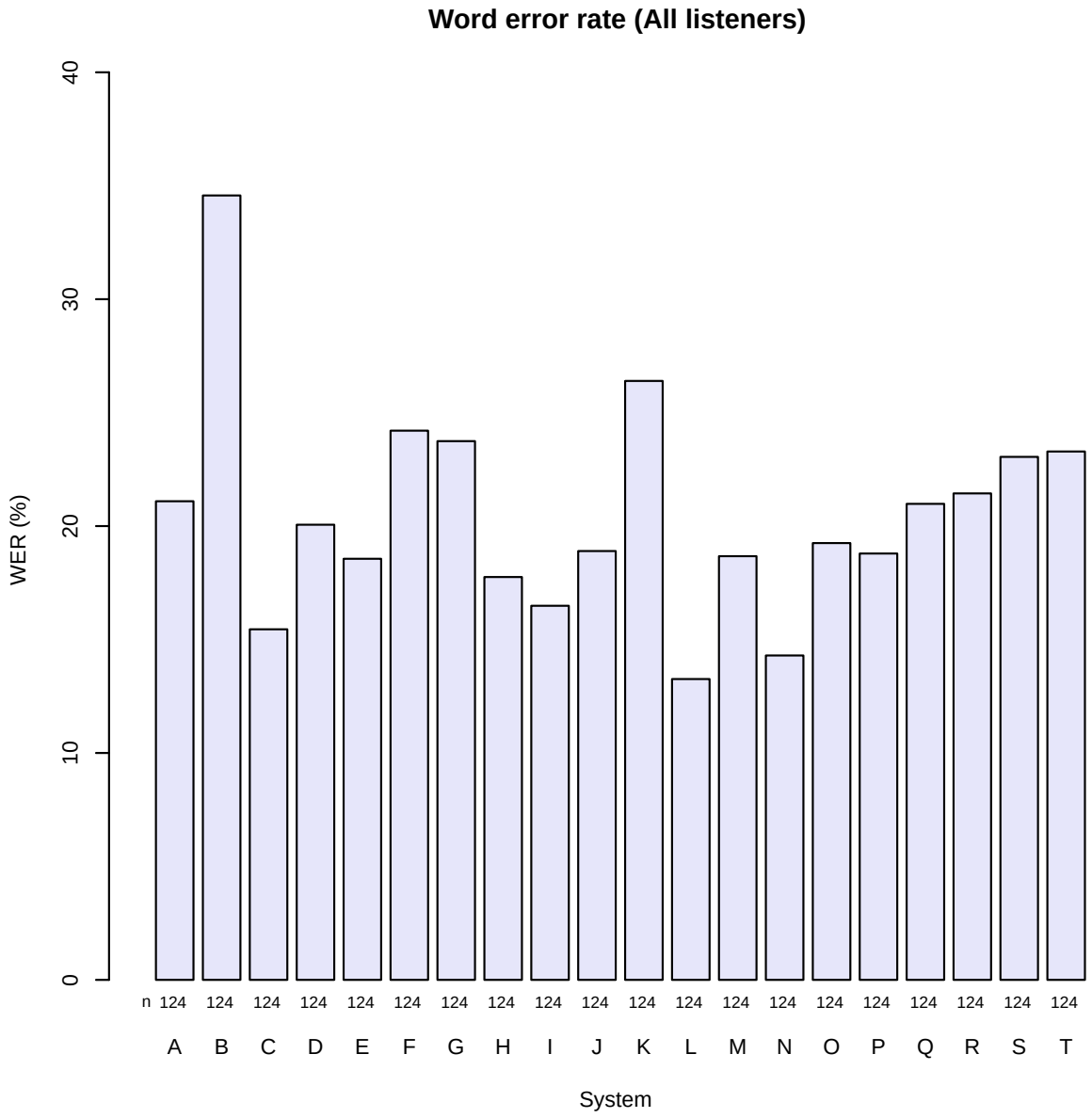


Fig. 3. TTS listening test results: intelligibility. See Table VII for system details.

REFERENCES

[1] M. Ostendorf and I. Bulyko, "The impact of speech recognition on speech synthesis," in *Proc. IEEE Workshop on Speech Synthesis*, Santa Monica, USA, Sep. 2002, pp. 99–106.

[2] M. Gales and S. Young, "The application of hidden Markov models in speech recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2007.

[3] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, doi:10.1016/j.specom.2009.04.004 2009.

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

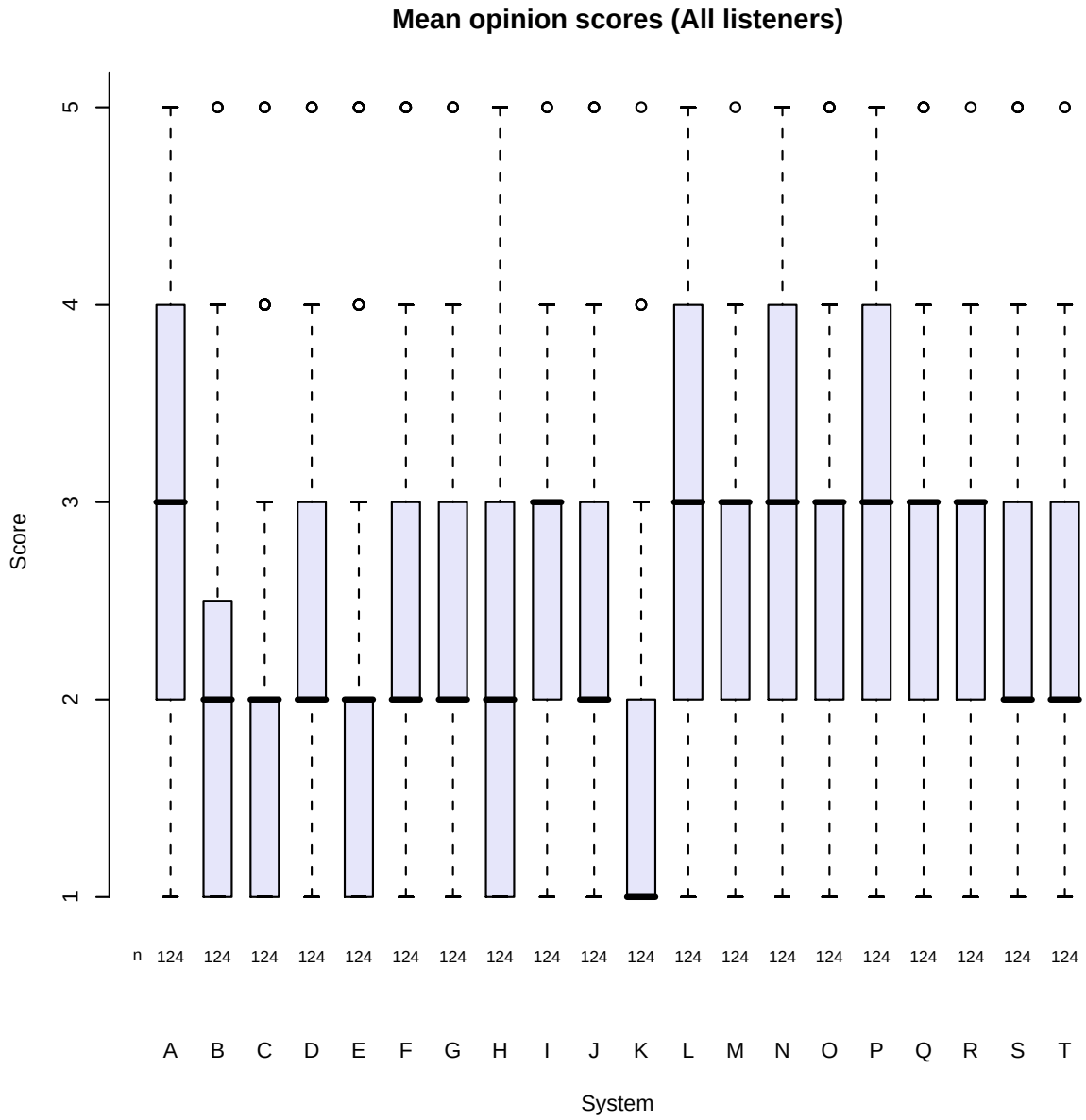


Fig. 4. TTS listening test results: naturalness. See Table VII for system details.

[5] M. Gibson, "Two-pass decision tree construction for unsupervised adaptation of HMM-based synthesis models," in *Proc. Interspeech*, Brighton, UK, September 2009, pp. 1791–1794.

[6] H. Liang, J. Dines, and L. Saheer, "A comparison of supervised and unsupervised cross-lingual speaker adaptation approaches for HMM-based speech synthesis," in *Proc. ICASSP*, Dallas, USA, 2010.

[7] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML*, Williamstown, USA, 2001, pp. 282–289.

[8] P. Brown, "The acoustic-modelling problem in automatic speech recognition," Ph.D. dissertation, Carneggie-Mellon University, 1987.

[9] D. Povey and P. C. Woodland, "Minimum Phone Error and I-Smoothing for improved discriminative training," in *Proc. ICASSP*, Orlando, USA, 2002.

[10] Y.-J. Wu and R.-H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Proc. ICASSP*, Toulouse, France,

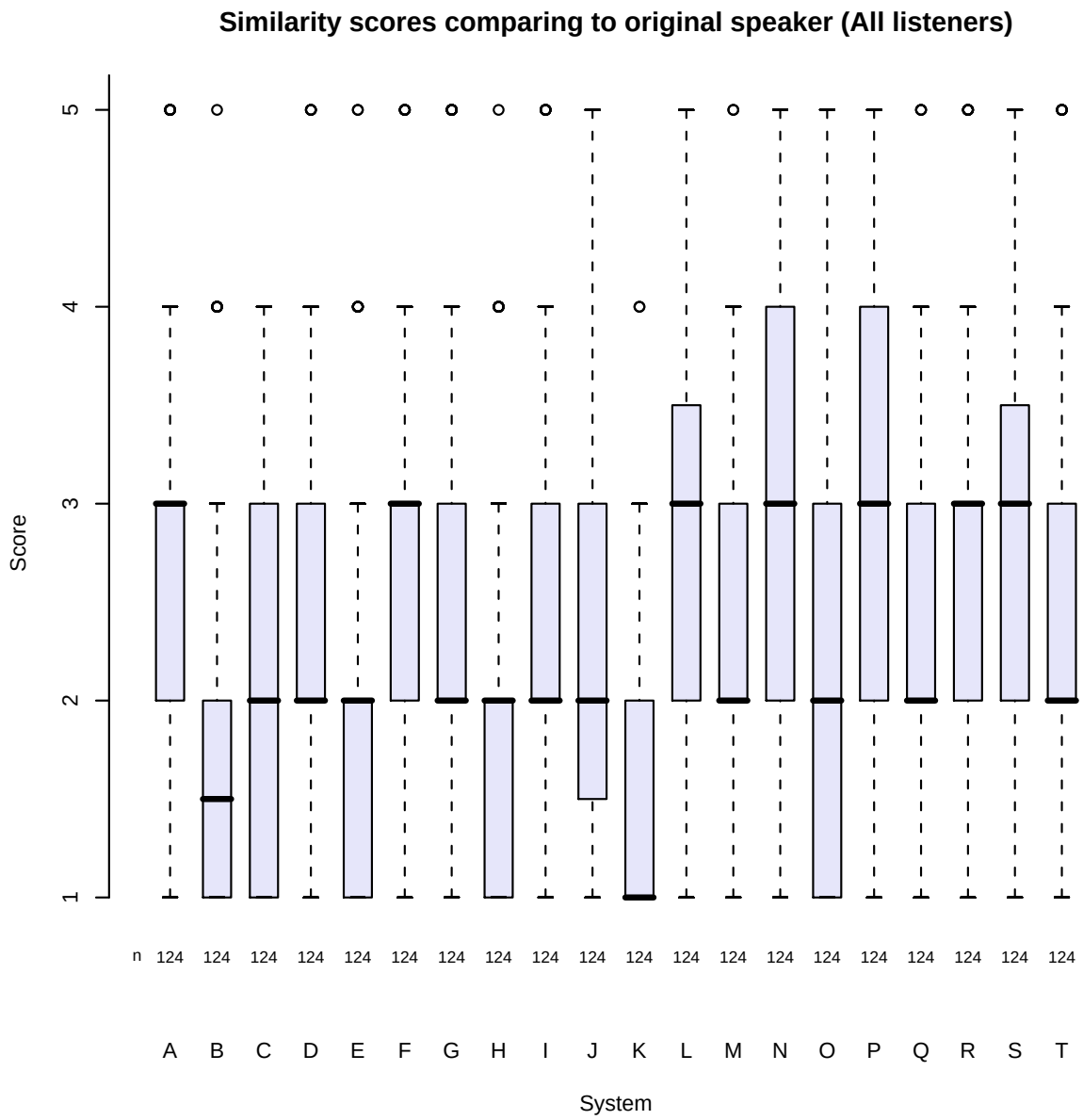


Fig. 5. TTS listening test results: similarity to original speaker. See Table VII for system details.

2006.

[11] L. R. Rabiner, "A tutorial on hidden Markov models and selected appications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[12] A. Falaschi, M. Giustiniani, and M. Verola, "A hidden Markov model approach to speech synthesis," in *Proc. Eurospeech*, Paris, France, 1989, pp. 187–190.

[13] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. on Acoustics, Speech, & Signal Process.*, vol. 29, pp. 254–272, April 1981.

[14] S.-Z. Yu, "Hidden semi-Markov models," *Artificial Intelligence*, vol. 174, no. 2, pp. 215–243, February 2009.

[15] K. Tokuda, H. Zen, and T. Kitamura, "Trajectory modeling based on HMMs with explicit relationship between static and dynamic features," in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 865–868.

[16] L. Deng, "A generalised hidden Markov model with state-conditioned trend functions of time for the speech signal," *Signal Processing*, vol. 27, pp. 65–78, April 1992.

[17] J. Dines, S. Sridharan, and M. Moody, "Trainable speech synthesis with trended hidden Markov models," in *Proc. ICASSP*, Salt Lake City, USA, 2001.

[18] C. Wellekens, "Explicit time correlation in hidden Markov models for speech recognition," in *Proc. ICASSP*, vol. 12, Dallas, USA, 1987.

[19] M. Shannon and W. Byrne, "Autoregressive HMMs for speech synthesis," in *Proc. Interspeech*, Brighton, UK, 2009.

[20] L. Deng and J. Ma, "A statistical coarticulatory model for the hidden vocal-tract-resonance dynamics," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 1499–1502.

[21] B. Mesot and D. Barber, "Switching linear dynamical systems for noise robust speech recognition," *IEEE Trans. Audio, Speech & Language Process*, vol. 15, no. 6, pp. 1850–1858, August 2007.

[22] M. Shannon and W. Byrne, "Autoregressive clustering for HMM speech synthesis," in *Proc. Interspeech*, Makuhari, Japan, 2010.

[23] Y. Bengio, "Learning deep architectures for AI," Université de Montréal, Montreal, Canada, Tech. Rep. 1312, 2007.

[24] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[25] A.-R. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *Proc. NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, Whistler, Canada, 2009.

[26] Y. Bengio, O. Delalleau, and C. Simard, "Decision trees do not generalize to new variations," Université de Montréal, Montreal, Canada, Tech. Rep. 1304, 2006.

[27] Y. Nankaku, K. Nakamura, H. Zen, T. Toda, and K. Tokuda, "Acoustic modelling with contextual additive structure for HMM-based speech recognition," in *Proc. ICASSP*, Las Vegas, USA, 2008.

[28] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, 3rd ed., Cambridge University Engineering Department, UK, December 2006.

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

[30] 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*, 1999, pp. 2347–2350.

[31] 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. Prentice Hall, 2004.

[32] 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.

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

[34] T. Hain, "Hidden model sequence models for automatic speech recognition," Ph.D. dissertation, Cambridge University, 2001.

[35] S. Dupont, H. Bourlard, O. Deroo, V. Fontaine, and J.-M. Boite, "Hybrid HMM-ANN systems for training independent tasks: Experiments on phonebook and related improvements," in *Proc. ICASSP*, Munich, Germany, April 1997, pp. 1767–1770.

[36] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[37] K. Koishida, G. Hirabayashi, K. Tokuda, and T. Kobayashi, "Mel-generalized cepstral analysis —a unified approach to speech spectral estimation," in *Proc. ICSLP*, vol. 3, Yokohama, Japan, September 1994, pp. 1043–1046.

[38] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "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.

[39] G. Garau and S. Renals, "Combining spectral representations for large vocabulary continuous speech recognition," *IEEE Trans. Speech, Audio & Language Process*, vol. 16, no. 3, pp. 508–518, March 2008.

[40] 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, February 2007.

[41] T. Irino, Y. Minami, T. Nakatani, M. Tsuzaki, and H. Tagawa, "Evaluation of a speech recognition / generation method based on HMM and STRAIGHT," in *Proc. ICSLP*, Denver, USA, 2002, pp. 2545–2548.

[42] J. J. Odell, "The use of context in large vocabulary continuous speech recognition," Ph.D. dissertation, Queens College, University of Cambridge, 1995.

[43] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. Eurospeech*, vol. 1, Rhodes, Greece, 1997, pp. 99–102.

[44] K. Prahallad, A. W. Black, and R. Mosur, "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis," in *Proc. ICASSP*, Toulouse, France, 2006, pp. 853–856.

[45] 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.

[46] 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.

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

[48] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 291–298, Apr. 1994.

[49] V. Digalakis and L. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Trans. Speech Audio Process.*, vol. 4, pp. 294–300, Jul. 1996.

[50] C. Leggetter and P. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression," in *Proc. ARPA Spoken Language Technology Workshop*. Morgan Kaufmann, 1995, pp. 104–109.

[51] O. Siohan, T. Myrvoll, and C.-H. Lee, "Structural maximum a posteriori linear regression for fast hmm adaptation," *Computer, Speech and Language*, vol. 16, no. 1, pp. 5–24, January 2002.

[52] Y. Nakano, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis," in *Proc. ICSLP 2006*, Sep. 2006, pp. 2286–2289.

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

[54] K. Oura, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "Hidden semi-Markov model based speech recognition system using weighted finite-state transducer," in *Proc. ICASSP*, Toulouse, France, May 2006, pp. 33–36.

[55] J. Dines, L. Saheer, and H. Liang, "Speech recognition with speech synthesis models by marginalising over decision tree leaves," in *Proc. Interspeech*, Brighton, UK, September 2009, pp. 1395–1398.

[56] 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 Workshop*, September 2008.

[57] 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.

[58] K. Tokuda, T. K. T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, Istanbul, Turkey, 2000, pp. 1315–1318.

[59] 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.

[60] D. Pallet, "DARPA February 1992 pilot corpus CSR "dry run" benchmark test results," in *Proceedings of the workshop on Speech and Natural Language*, Harriman, USA, February 1992, pp. 382–386.

[61] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, K. Oura, K. Tokuda, R. Karhila, and M. Kurimo, "Thousands of voices for HMM-based speech synthesis," in *Proc. Interspeech*, Brighton, UK, September 2009, pp. 420–423.

[62] ——, "Thousands of voices for HMM-based speech synthesis – analysis and application of TTS systems built on various ASR corpora," *IEEE Trans. Speech, Audio & Language Process*, vol. 18, no. 5, pp. 984–1004, July 2010.

[63] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. ICASSP*, vol. 1, Montreal, Canada, May 1994, pp. 409–412.

[64] A. Gray Jr. and J. Markel, "Distance measures for speech processing," *IEEE Trans. on Acoustics, Speech, & Signal Process.*, vol. 24, no. 5, pp. 380–391, Oct. 1976.

[65] T. P. Barnwell III, "Correlation analysis of subjective and objective measures for speech quality," in *ICASSP*, 1980, pp. 706–709.

[66] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[67] T. Kitamura, S. Imai, C. Furuichi, and T. Kobayashi, "Speech analysis-synthesis system and quality of synthesized speech using mel-cepstrum," *Electronics and Communications in Japan (Part I: Communications)*, vol. 69, no. 10, pp. 47–54, 1986, (in Japanese).

[68] T. Fukada, K. Tokuda, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP 1992*, San Francisco, CA, 1992, pp. 137–140.

[69] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, 1993.*, vol. 1, May 1993, pp. 125–128 vol.1.

[70] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

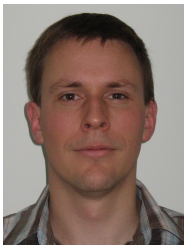
[71] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *Proc. Blizzard Challenge Workshop*, Brisbane, Australia, September 2008.

[72] "The CMU pronouncing dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

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

[74] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 639–643, October 1994.

[75] J. Dines, J. Yamagishi, and S. King, "Measuring the gap between HMM-based ASR and TTS," in *Proc. Interspeech*, Brighton, UK, September 2009, pp. 1391–1394.



John Dines (M'99) graduated with first class honours in Electrical and Electronic Engineering from University of Southern Queensland in 1998 and received the Ph.D. degree from the Queensland University of Technology in 2003 with the thesis: "Model based trainable speech synthesis and its applications". Since 2003 he has been employed at the Idiap Research Institute, 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. He is a member of IEEE and a reviewer for IEEE Signal Processing Letters and IEEE Transactions on Audio, Speech and Language Processing.

Junichi Yamagishi received the B.E. degree in computer science, M.E. and Ph.D. degrees in information processing from Tokyo Institute of Technology, Tokyo, Japan, in 2002, 2003, and 2006, respectively. He 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 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, 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). He has over 50 refereed publications. His research interests include speech synthesis, speech analysis, and speech recognition. He is a member of IEEE, ISCA, IEICE, and ASJ.

Simon King (M'95–SM'08) received the M.A.(Cantab) degree in Engineering and the M.Phil. degree in Computer Speech and Language Processing from the University of Cambridge, Cambridge, UK in 1992 and 1993 respectively and the Ph.D. degree in speech recognition from the University of Edinburgh in 1998. He has been involved in speech technology since 1992, and has been with the Centre for Speech Technology Research, University of Edinburgh, since 1993. He is a Reader in Linguistics and English Language and an EPSRC Advanced Research Fellow. His interests include concatenative and HMM-based speech synthesis, speech recognition and signal processing, with a focus on using speech production knowledge to solve speech processing problems. He is a member of ISCA, serves on the steering committee for SynSIG (the special interest group on speech synthesis) and co-organises the Blizzard Challenge. He is member of the IEEE and an associate editor of IEEE Transactions on Audio, Speech and Language Processing.

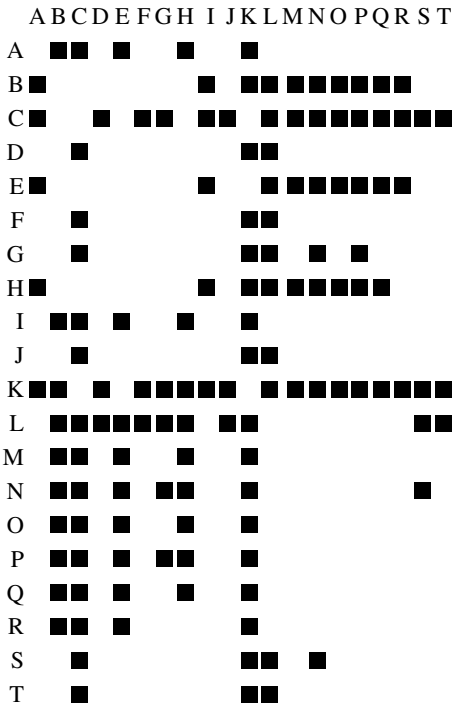


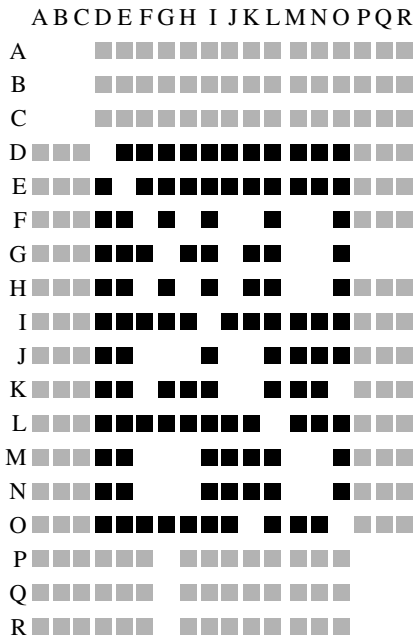
TABLE VIII

SIGNIFICANT DIFFERENCES IN NATURALNESS: RESULTS OF PAIRWISE WILCOXON SIGNED RANK TESTS BETWEEN SYSTEMS' MEAN OPINION SCORES. ■ INDICATES A SIGNIFICANT DIFFERENCE BETWEEN A PAIR OF SYSTEMS. SEE TABLE VII FOR SYSTEM DETAILS.

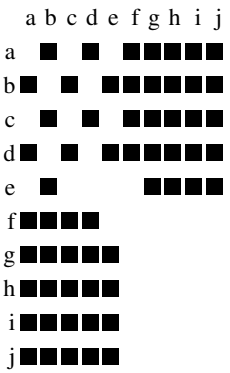
	Test	Pronunciation		Spectral analysis		Decision tree		Adaptation			WER
Index	set	Lexicon	Phone set	Method	Order	Structure	Stopping	Algorithm	Supervised	First-pass	(%)
A	H2 (P0)	**CMU**	**CMU**	PLP	13	phonetic	HTK	CMLLR	N	SI 5k-bg	6.4
B	H2 (P0)	**Unisyn**	**GAM**	PLP	13	phonetic	HTK	CMLLR	N	SI 5k-bg	6.6
C	H2 (P0)	**Unisyn**	**Arpabet**	PLP	13	phonetic	HTK	CMLLR	N	SI 5k-bg	6.1
D	H2 (P0)	Unisyn	GAM	**PLP**	**13**	phonetic	MDL	CMLLR	N	SI 5k-bg	6.8
E	H2 (P0)	Unisyn	GAM	PLP	**25**	phonetic	MDL	CMLLR	N	SI 5k-bg	8.1
F	H2 (P0)	Unisyn	GAM	PLP	**40**	phonetic	MDL	CMLLR	N	SI 5k-bg	11.9
G	H2 (P0)	Unisyn	GAM	**MCEP**	**13**	phonetic	MDL	CMLLR	N	SI 5k-bg	9.4
H	H2 (P0)	Unisyn	GAM	MCEP	**25**	phonetic	MDL	CMLLR	N	SI 5k-bg	10.9
I	H2 (P0)	Unisyn	GAM	MCEP	**40**	phonetic	MDL	CMLLR	N	SI 5k-bg	19.1
J	H2 (P0)	Unisyn	GAM	**MCEP+**	**13**	phonetic	MDL	CMLLR	N	SI 5k-bg	11.4
K	H2 (P0)	Unisyn	GAM	MCEP+	**25**	phonetic	MDL	CMLLR	N	SI 5k-bg	12.8
L	H2 (P0)	Unisyn	GAM	MCEP+	**40**	phonetic	MDL	CMLLR	N	SI 5k-bg	16.0
M	H2 (P0)	Unisyn	GAM	**MGCEP+**	**13**	phonetic	MDL	CMLLR	N	SI 5k-bg	10.3
N	H2 (P0)	Unisyn	GAM	MGCEP+	**25**	phonetic	MDL	CMLLR	N	SI 5k-bg	10.2
O	H2 (P0)	Unisyn	GAM	MGCEP+	**40**	phonetic	MDL	CMLLR	N	SI 5k-bg	13.6
G	H2 (P0)	Unisyn	GAM	MCEP	13	phonetic	MDL	CMLLR	N	SI 5k-bg	9.4
P	H2 (P0)	Unisyn	GAM	MCEP	13	**phonetic**	HTK	CMLLR	N	SI 5k-bg	9.4
Q	H2 (P0)	Unisyn	GAM	MCEP	13	shared	MDL	CMLLR	N	SI 5k-bg	9.2
R	H2 (P0)	Unisyn	GAM	MCEP	13	**shared**	HTK	CMLLR	N	SI 5k-bg	9.4
a	S4 (C3)	Unisyn	GAM	MCEP	13	phonetic	MDL	**MLLR**	Y	–	11.5
b	S4 (C3)	Unisyn	GAM	MCEP	13	phonetic	MDL	**CMLLR**	Y	–	13.2
c	S4 (C3)	Unisyn	GAM	MCEP	13	phonetic	MDL	**SMAPLR**	Y	–	11.5
d	S4 (C3)	Unisyn	GAM	MCEP	13	phonetic	MDL	**CSMAPLR**	Y	–	13.0
e	S4 (C3)	Unisyn	GAM	MCEP	13	phonetic	MDL	**SMAPLR+MAP**	Y	–	11.5
f	S4 (C3)	Unisyn	GAM	MCEP	13	phonetic	MDL	**CSMAPLR+MAP**	Y	–	13.9
g	S4 (–)	Unisyn	GAM	MCEP	13	phonetic	MDL	CSMAPLR+MAP	**N**	**SI 5k-bg**	14.3
h	S4 (–)	Unisyn	GAM	MCEP	13	phonetic	MDL	CSMAPLR+MAP	**N**	**SAT 5k-bg**	14.4
i	S4 (–)	Unisyn	GAM	MCEP	13	phonetic	MDL	CSMAPLR+MAP	**N**	**SI 20k-bg**	14.3
j	S4 (–)	Unisyn	GAM	MCEP	13	phonetic	MDL	CSMAPLR+MAP	**N**	**SAT 20k-bg**	14.2

TABLE IX

THE 28 ASR SYSTEMS THAT WERE EVALUATED. BOLD FACE IS USED TO HIGHLIGHT THE SETTING(S) BEING VARIED IN EACH SUBSET OF RESULTS. MCEP+, MGCEP+ ARE ABBREVIATIONS OF STRAIGHT+MCEP AND STRAIGHT+MGCEP RESPECTIVELY.



(a) H2 task



(b) S4 task

TABLE X

SIGNIFICANT DIFFERENCES IN ASR SYSTEMS AT 95% CONFIDENCE. ■ INDICATES A SIGNIFICANT DIFFERENCE BETWEEN A PAIR OF SYSTEMS, ■ INDICATES SIGNIFICANCE TEST WAS NOT RUN ON GIVEN PAIR. SEE TABLE IX FOR SYSTEM DETAILS.