Measuring the gap between HMM-Based ASR and TTS

J. Dines, J. Yamagishi, and S. King. Measuring the gap between HMM-based ASR and TTS. IEEE Selected Topics in Signal Processing, vol 4.6.

Measuring the gap between HMM-based ASR and TTS

John Dines[1]*, Junichi Yamagishi*[2]*, Simon King*[2]

[1]Idiap Research Institute, Martigny, Switzerland
[2] The Centre for Speech Technology Research, University of Edinburgh, Edinburgh, U.K.
john.dines@idiap.ch, jyamagis@inf.ed.ac.uk, simon.king@ed.ac.uk

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 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 and HMM topology. 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

1. Introduction

Over the last decade speech recognition and speech synthesis technologies have shown a convergence towards statistical parametric approaches . In the EMIME[1] project we are exploiting this convergence in the context of speech-to-speech translation (ST). More specifically, we are using HMM-based automatic speech recognition (ASR) and text-to-speech (TTS) in order to achieve two goals in ST: firstly, the ability to efficiently adapt system to the user's voice and, secondly, in the context of a mobile application, we wish to benefit from the parsimonious nature of such approaches.

The use of unified models in ST represents a particularly attractive paradigm since it provides a natural mechanism for speaker-adaptive synthesis by employing the same speaker dependent transforms learned from ASR, while offering further efficiency with respect to computation and memory (see for eg. [1]). 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 an empirical study of ASR and TTS systems . 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 paper is primarily concerned with 'measuring the gap' between ASR and TTS.

The paper is organised as follows: Section 2 briefly presents statistical modelling for ASR and TTS, outlining the major differences between the two. Section 3 describes our methodology and Section 4 details our empirical studies and analysis in measuring the the gap between ASR and TTS systems. Finally in Section 5 we present our conclusions.

2. HMM-based ASR and TTS

The hidden Markov model has been dominant paradigm of 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: scalability to large scale tasks; desirable generalisation properties; powerful adaptation framework; and parsimony with respect to the quantity 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 [2], a trend that is also continuing in TTS with HTS [3]. 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[4].

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 1 shows typical configurations of HMM-based ASR and TTS systems. For further details please refer to [2, 5].

2.1. Lexicon and phone set

The lexicon describes the set of words known by the ASR/TTS system and their respective pronunciations. In TTS we may also generate pronunciations that lie outside of the lexicon us-

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

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

Table 1: Configuration of HMM-based ASR and TTS systems.

ing letter-to-sound (LTS) prediction. 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 self-evident that in both cases phone sequences produced by the lexicon should have good correlation with acoustic data.

2.2. Feature extraction

Typically, feature extraction techniques differ between speech recognition and speech synthesis. In speech 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) . TTS features are normally based on variations of the mel-generalised cepstrum analysis and normally incorporate STRAIGHT pitch-adaptive spectral analysis.

2.3. 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. We may also consider parameter smoothing and parameter tying techniques, such as decision tree state tying, as being concerned with model topology. In ASR, it is typical to employ 3-state left-right HMM with phonetic decision trees, whereas in TTS, 5-state left-right with a shared decision tree per state are used.

3. Methodology

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 1. In these experiments we are not considering such fundamental differences as duration or context modelling – these being the subject of more focused research.

3.1. Experimental setup

We built the ASR and TTS systems based on the HTS entry to the 2007 Blizzard Challenge [5]. Thus, ASR and TTS models are trained using maximum likelihood speaker-adaptive training (ML-SAT). We also train speaker independent models (ML-SI) for first-pass ASR decoding. As far as possible, the variables for experimentation (e.g. training and test sets, speech features, and so on) are shared between both ASR and TTS systems. Training data comprised the Wall Street Journal (WSJ0) short term speaker training data (SI84) which includes 7240 recordings made by 84 speakers [6]. While there are obvious disadvantages to training speech synthesis from an ASR corpus, there is the advantage in that it grants access to larger quantities of data and speakers compared to that which is available from current TTS corpora. For further details see [7].

3.2. ASR evaluation

For the evaluation of ASR we use the primary condition (P0) of the 5k vocabulary hub task (H2) of the November 93 CSR evaluations. The decoding employs the 5k closed bigram language model distributed with the corpus and is carried out using speaker independent models for the first pass and SAT trained models in the second pass. The word error rate (WER) metric to evaluate ASR system performance.

3.3. TTS evaluation

For the evaluation of TTS we also used the November 1993 CSR H2 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 only for a sub-

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	5.56	198	15.7
Unisyn	Arpabet (45)	6.1	5.60	198	16.3

Table 2: Comparisons of lexica for ASR and TTS.

set of systems, and for a single target speaker, '4oa'. Objective measures have been used for all systems and all the target speakers.

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), which is only calculated during periods of speech activity. 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 Blizzard Challenge 2008 [5, 8]. 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 are conducted via a standard web browser with a total of 124 native English speakers participating in these tests.

4. Results and analysis

4.1. Comparison of phone set and lexicon

The CMU lexicon [9] was used in the baseline ASR system and the Unisyn lexicon [10] 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 Arpabet-like set 45 phonemes was also evaluated. The results of lexicon evaluations are shown in Table 2. 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. Observations for TTS are to the contrary of ASR with the Unisyn lexicon giving better objective measures 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.

4.2. 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. 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. We do not consider features for $\log F_0$ or aperiodicity measures (bndap) in ASR experiments. The results of these comparisons are shown in Table 3.

First of all, we see that conventional ASR features (PLP) 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. By contrast, there is no straight forward method for synthesising features from filter-bank based features, though alternative techniques have been investigated in the context of distributed speech recognition that would deserve consideration for TTS [11]. Of all of the MCEP-based features, the STRAIGHT+MGCEP features provide the best performance on average for ASR, which is consistent with results reported in the literature. For TTS, there is little to separate the different feature analysis methods for such a speaker adaptive system.

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. Results of particular interest were obtained with the MGCEP+STRAIGHT features at an analysis order of 25, which show minimal degradation to performance of ASR and TTS when compared, respectively, to lower and higher analysis orders.

When considering the most likely explanations for this behaviour we recall that lower order cepstra are generally considered 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. 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).

There are several possible approaches that may be investigated to alleviate the feature dimensionality problem. Perhaps the simplest approach would be to place high and low order cepstra in separate streams where the stream weighting for higher order cepstra would be zero for ASR. Similarly, linear transforms may provide means for dimensionality reduction while minimising impact on ASR and TTS. Lastly, the use of features for source modelling ($\log F_0$ and bndap) may also be considered for ASR.

4.3. Comparison of model topology

We conducted experiments with respect to HMM topology by comparing different state tying schemes, where the ASR base-

Feature		ASR			TTS	
Type	Order	WER (%)			WER	MOS
		All	Male	Female	(%)	
PLP	13	7.0	8.4	5.6	–	–
	25	8.0	9.2	7.2	–	–
	40	10.6	11.1	10.0	–	–
STRAIGHT+MCEP	13	11.7	13.6	9.1	15	1.9
	25	12.4	13.2	10.7	20	2.4
	40	21.7	21.5	22.0	21	2.7
STRAIGHT+MGCEP	13	10.4	12.8	8.0	19	2.0
	25	10.4	12.5	8.3	24	2.5
	40	15.1	18.3	11.8	24	2.3
STRAIGHT+MGLSP	13	–	–	–	18	2.0
	25	–	–	–	16	2.7
	40	–	–	–	19	2.5

Table 3: Comparisons of features for ASR and TTS.

line uses phonetic decision trees (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 state) with MDL criterion to control tree growth. These two configurations offer their own pros and cons, 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 4 shows the results for these experiments.

An unexpected result for the ASR experiments revealed that the shared decision tree yielded equal or 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 models. To what extent these two factors affect system performance must depend on the training data, phone set and lexicon.

The TTS results show that, contrary to ASR, 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 light of these results, it would appear that shared decision trees may be suitable for both ASR and TTS, though this may be tied to the phone set being used. An alternative configuration would take on a hybrid approach, in which spectrum feature streams would use phonetic decision trees and $\log F_0$ feature streams would use shared decision trees.

5. 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, in general, many of the techniques used in ASR and TTS can not be simply applied to their respective other without negative consequences. In particular, we have identified that feature extraction and feature order have the most detrimental impact on ASR and TTS performance. It is clear that further research will need to concentrate on these two aspects if we wish to make significant inroads to bridging the gap between ASR and TTS.

Tree type	Criteria	Threshold		ASR	TTS		
		Likelihood	Occupancy	WER	MCD	RMSE	V/UV
		(TB)	(RO)	(%)		$\log F_0$	error
Phonetic	ML	450	200	9.4	–	–	–
	MDL	–	–	9.4	5.66	447	15.9
Shared	ML	300	200	9.4	–	–	–
	MDL	–	–	9.2	5.56	198	15.7

Table 4: Comparisons of state-tying for ASR and TTS. The ASR systems use 13-dimensional MCEP features and the TTS systems use 40-dimension STRAIGHT+MCEP features.

6. Acknowledgements

The research leading to these results was partly funded from the European Communitys 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).

7. References

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

[2] S. Young et. al., *The HTK Book*, 3rd ed., Cambridge University Engineering Department, UK, December 2006.

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

[4] T. Yoshimura et. al., "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech*, 1999, pp. 2347–2350.

[5] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "HTS-2007 system for the Blizzard challenge 2007," in *Proc. of Blizzard Challenge 2007 workshop)*, Bonn, Germany, August 2007.

[6] 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.

[7] J. Yamagishi et. al., "Thousands of voices for HMM-based speech synthesis," in *Proc. Interspeech 2009*, Brighton, UK, September 2009.

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

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

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

[11] B. Milner and X. Shao, "Clean speech reconstruction from MFCC vectors and fundamental frequency using an integrated front-end," *Speech Communication*, vol. 48, no. 6, pp. 697–715, 2006.