Thousands of voices for HMM-based speech synthesis

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:

Publisher Rights Statement:

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Abstract

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

Index Terms: speech synthesis, HMMs, speaker adaptation

1. Introduction

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

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

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

2. HTS voices trained on ASR corpora

2.1. Framework of TTS systems

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

2.2. ASR speech databases used for TTS systems

In conventional speech synthesis research, phonetically-balanced speech databases are typically used. A phonetically-balanced dataset (e.g., complete diphone coverage) is required for each individual speaker, since conventional systems are speaker-dependent. In multi-speaker sets of speech synthesis data (e.g., CMU-ARCTIC1), it is common for the same set of phonetically-balanced sentences to be re-used for each speaker. Therefore, pooling the data from multiple speakers does not always significantly increase phonetic coverage. Compared to this, the sentences chosen for ASR corpora tend to be designed to achieve phonetic balance across multiple speakers, or simply chosen randomly. Therefore, phonetic coverage increases with the number of speakers. However, each individual speaker typically records a very limited number of utterances (e.g., fewer than 100). Building TTS voices from these ASR corpora is in itself a new challenge.

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

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

1A free database for speech synthesis, http://festvox.org/cmua_rctic/
Table 1: Triphone coverage of ASR and TTS corpora

<table>
<thead>
<tr>
<th>Name</th>
<th>triphones/speaker</th>
<th>triphones/corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMU-ARCTIC</td>
<td>10041</td>
<td>10708</td>
</tr>
<tr>
<td>WSJ0 (short/SI-84)</td>
<td>3287</td>
<td>18577</td>
</tr>
<tr>
<td>WSJ0+1 (short/SI-284)</td>
<td>4220</td>
<td>23776</td>
</tr>
<tr>
<td>WSJCAM0 (total)</td>
<td>3036</td>
<td>23534</td>
</tr>
<tr>
<td>RM (nd, total)</td>
<td>1091</td>
<td>7162</td>
</tr>
</tbody>
</table>

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

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

2.3. Front-end processing

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

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

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

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

2.5. Demonstration of the HTS voices

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

2.6. Geographical representation and online demo

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

2.7. Multidimensional scaling of male speakers included in
WSJ0 corpus

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

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

3. Evaluation

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

3.1. Average voice model training data

We built two kinds of average voice model. The first was built
using 50 utterances per training speaker (“condition 1”). If a
speaker has more than 50 utterances, a subset of 50 was cho-

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

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

3.2. Speaker-dependent model training data

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

3.3. Objective evaluation

Table 2 shows the objective measures for each system. From
the results for speaker 001, we can confirm that the speaker-
adaptive systems using all available average voice model train-
ing data (“condition 2”) outperform the speaker-adaptive sys-
\(^2\)Currently the interactive mode supports English and Spanish only.
For other languages this only provides pre-synthesised examples, but
we plan to add an interactive type-in text-to-speech feature in the near
future.

..., and 2000 randomly chosen sentences
...ative systems start to become
better than speaker-adaptive systems. This result is consistent
with previous results.

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

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

Acknowledgements

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

5. References


