From Pre-recorded Prompts to Corporate Voices: On the Migration of Interactive Voice Response Applications

V. Fischer, S. Kunzmann

IBM Deutschland Entwicklung GmbH,
European Voice Technology Development & Services
Schönaicher Straße 12, D-71032 Böblingen, Germany
vfischer@de.ibm.com

10.21437/Interspeech.2006-405

Abstract

This paper describes our efforts towards the creation of corporate synthetic voices from low quality speech data, as it can typically be found on many Interactive Voice Response (IVR) units. In doing so, we first touch on several normalization techniques that aim on a better support of a highly automated voice construction process. Subsequently, we describe methods for the creation of *enriched* corporate voices which integrate speech recordings from different speakers in order to overcome problems arising from limited domain training data.

Experiments are described which demonstrate the feasibility of the approach by comparing it to a less flexible solution that uses pre-recorded prompts in combination with a large footprint standard concatenative synthesizer. Results show that the enriched voices clearly outperform those voices build solely from IVR data, while achieving almost the same overall rating as the pre-recorded prompts solution.

Index Terms: speech synthesis, IVR migration, corporate voices, limited domain synthesis, pooled speaker synthesis.

1. Introduction

Due to major progress in the field of concatenative speech synthesis, cf. e.g. [1, 2, 3, 4], todays voice driven applications can make use of a variety of high quality synthetic voices for the creation of natural sounding speech output in many languages. However, both voice user interface developers who start to add personalization features to their application and service providers who wish to brand their application with a unique, immediately identifiable *corporate voice* create a still increasing demand for more synthetic voices.

While a better utilization of synergies between speech synthesis and automatic speech recognition (see, for example, [5]), can free text-to-speech developers from tedious manual work and thus help to significantly reduce efforts for the creation of synthetic voices, the migration of existing applications that make use of pre-recorded prompts into corporate voice applications remains a non-trivial task for several reasons:

- The recording of speech databases for text-to-speech systems requires a careful control of the acoustic environment as well as the thoughtful direction of the speaker. The construction of synthetic voices from speech databases that do not comply to TTS specific recording standards may require modifications to meanwhile widely automated voice building processes, or may result in low quality speech output otherwise.
- One reason why TTS systems still sound best if synthesizing text from domains included in the recording script, cf. [6], is the proper coverage of phonetic contexts, which results in less splices in the synthetic speech, and thus in fewer audible distortions. While almost achieving the quality of pre-recorded prompts when synthesizing application specific carrier phrases, synthetic voices constructed solely from limited domain data may therefore produce less than optimal speech output for unseen or variable text, like e.g. proper names.
- Re-recording or capturing of additional data can help to overcome these deficiencies [7], but may be prohibitive due to monetary reasons, or because the original speaker is no longer available.

The work presented in this paper deals with the issues mentioned above by providing a comparison of a cost-efficient solution — the use of pre-recorded audio together with a general purpose synthetic voice — with two domain specific corporate voices, one being created solely from the prompts used in the first scenario, the other being enriched with a small amount of additional data *from a different speaker*.

For that purpose, we start with a brief description of the IBM trainable speech synthesis system in the next sec-

tion, before we discuss the construction of the two corporate voices in Section 3. Finally, Section 4 provides some experimental evaluations of the three different scenarios, and Section 5 concludes with a summary and an outlook on some future work.

2. System Overview

Basic principles and most recent version of the IBM trainable text-to-speech system are documented in some detail in [2, 8]. The system uses subphoneme-sized speech segments as synthesis units, which correspond to context-dependent Hidden-Markov-Model states that are identified by decision tree growing during the construction of the system. Decision trees are also used for the prediction of speaker dependent phone durations and pitch contours for each syllable to be synthesized.

During runtime, an independent rule-based front end is employed for text normalization, text-to-phone conversion, and phrase boundary generation. Preprocessed phrases are passed to the back-end that employs a Viterbi beam-search to generate the synthetic speech.

Slightly varying across languages and voices, the training of the system is based on a script of roughly 10K sentences (approx. 15 hours of speech, including silence) that were recorded by a professional speaker in a professional recording studio. The script includes a section of 1400-2000 sentences that were designed to achieve optimal diphone or triphone coverage, a mixture of newspaper articles and emails about different topics, weather forecasts, proper names (cities, persons, company names), digits and natural numbers, dates and times, and a variety of prompts that are related to various popular voice driven applications (e.g. air travel information and finance).

Among the mandatory steps in the construction of the synthesizer are the training of above mentioned decision trees for duration and pitch prediction, and the creation of the acoustic unit inventory from speaker dependent, pitch-synchronous alignments. For that purpose, a wavelet transform based algorithm for the extraction of pitch marks is employed [9], which has meanwhile replaced methods based on laryngograph recordings of the speaker's glottis signal.

Optional training steps include the creation of speaker specific runtime dictionaries for the reconciling of pronunciation differences between the synthesizers front-end and the back-end database [10], and a data driven approach to segment pre-selection [11].

3. Corporate Voices from IVR Data

Based on experiences gathered during the creation of corporate voices for several domains and languages, in this section we describe extensions to the training procedure outlined above that turned out to be beneficial when dealing with speech material of fairly low quality — sampled at 8 kHz and quantized with 8 bit — as it can be found on typical IVRs. Also described in this section are methods that aim on a better integration of data from different speakers, which can help to overcome problems with limited domain training data, cf. Section 4.

A preparatory step not discussed in detail here is the creation of a speaker independent set of HMMs for the initial alignment of 8 kHz data (with linear quantization). For that purpose, we simply transformed our standard set of multilingual 22 kHz alignment models by running one iteration of the forward-backward training algorithm with two parallel data streams, namely a 22 kHz MFCC feature stream for the computation of alignments, and the corresponding 8 kHz feature stream for the estimation of new model parameters.

The so created HMMs were used with variable rejection thresholds for an alignment based identification of transcription errors and text normalization problems. While the latter occurred in only one application scenario with many uncommon abbreviations, the quality of available transcriptions turned out to be rather poor in most cases, if compared to carefully prepared TTS recording scripts.

The speech data itself was first converted to 16 bit linear PCM format, and approx. 500 msec of pre-recorded silence was added to the begin and end of each speech file in order to match the training conditions of the speaker independent initial HMMs. The preprocessed audio files were further normalized in two ways: First, we removed clipping and large differences in signal amplitude across audio files by scaling the RMS energy to an average value. Second, in order to run the wavelet transform based pitch mark extraction with a fixed set of parameters in an unattended batch mode, we normalized the signal polarity. For that purpose we first computed the polarity ϕ for each speech file by simply using

$$\phi = \begin{cases} 1 & \text{if} \quad E_p > E_n \\ -1 & \text{else} \end{cases}$$

$$E_p = \sum_{x(t)>0} x(t) \cdot x(t)$$

with E_p (and E_n) being the signal energy computed from all positive (negative) samples. Obtaining two sets of signal files with different polarity, the smaller set of files was inverted by setting $\hat{x}(t) = -x(t)$ for all samples $x(t)$ and all files.

For the integration of additional rich-context data from one or more of our standard TTS voices (henceforth: *auxiliary speakers*) we normalized RMS energy across databases, and also moved the characteristics of the auxiliary speakers' voice towards the *reference* or *target speaker*, i.e. the pre-

recorded prompts of the corporate speaker.

Different from [12], where a shift in the average pitch is used for the creation of pitch prediction models from pooled speaker data, in our work we focused on the adjustment of the auxiliary speakers' spectral envelope towards the target speaker. For that purpose we designed an FIR equalization filter that minimizes the difference between the auxiliary speakers' and the target speaker's long term spectrum. In order to avoid overfitting to a particular reference sentence, we used several of the target speaker's utterances for the estimation of a reference spectrum, and randomly divided the auxiliary speaker database into subsets that were equalized with different filters.

4. Experiments

The main goal of the experiments described in this section is an evaluation of our approach to the creation of *enriched* corporate voices, which on the one hand should help to improve the overall voice quality, but on the other hand should also preserve the identity of the corporate speaker's voice. The scenario is defined within a phone banking application, which consists of fixed prompts (like, e.g.: `"You can say 'Help' at any time"`) as well as of carrier phrases with embedded variable text (e.g.: `"You want to transfer` *`60 Euro and 25 Cent`* `to` *`John Smith`*`. Is this correct?"`); in the listening tests described below we used only sentences of the second type, with proper names as variables.

Following the training procedure outlined in Section 2 and using the extensions described in Section 3 for the creation of the (enriched) corporate voice, we created three different German male 8kHz voices whose characteristics are summarized as follows:

- SV: the IBM German male standard voice ("Dieter"), created out of approx. 6.5 hours of phonetically rich sentences plus approx. 7 hours of application specific data from various domains, cf. Section 2, all recorded at 48 kHz.
- CV: a German male corporate voice, created out of approx. 6800 prompts (7 hours of speech) from a phone banking application that were available to us as real IVR data, i.e. μ-law coded at 8 kHz sample rate.
- EV: an enriched of CV, created out of the data used for CV, plus the 6.5 hours of phonetically rich sentences used for the construction of SV. As described in Section 3, the only preprocessing steps that aimed on making the two datasets more similar was spectral equalization and energy normalization across datasets.

Computed from the automatically extracted pitch marks, Table 1 gives the average pitch for the 3 voices, showing also a smaller difference after moving SV towards CV by means of long term spectral equalization.

voice	SV (orig.)	SV (equal.)	CV	EV
avg. pitch	96	101	112	109

Table 1: Average pitch (in Hz) of IBM's standard voice (SV, before and after spectral equalization), corporate voice (CV), and enriched corporate voice (EV).

Being created from different amounts of data, CV and EV were designed to end up with approximately the same number of candidate speech segments for each sub-phonetic context, and all three voices use the run-time parameters of the German male standard voice (SV). Note that these settings give favor to long contiguous segments (as can be found in the carrier phrases), while putting a little disadvantage on the synthesis of unseen text, like, for example, the proper names in our scenario. Different from [12] we made no attempt to tweak the cost function towards the use of data from a particular speaker (in our case: the corporate speaker), but plan to do so in a future row of experiments.

A small listening test with — alas — only 9 participants (4 of them being non natives with different proficiency of German) was carried out in order to compare the voices. The items that had to be judged on a 5 point scale were as follows:

- *Media break*: How noticeable is the use of speech from two different speakers? Note that for this question a lower score is favorable.
- *Intelligibility*: How easy is it to understand the prompt? A particular focus should be on the audibility of the free variable.
- *Prosody*: How natural and appropriate is the intonation? A particular focus should be on the application specific carrier phrases.
- *Overall impression*: Taking into account the above (and probably other personal preferences), how well is the sample liked?

Table 2 gives the mean opinion score (MOS) results for the comparison of the three voices, with SV used only for the synthesis of variable text items in combination with prompts pre-recorded by the corporate speaker.

A result that is in accordance with [12] is the preference for a synthetic voice that was created from pooled data. However, we were a bit surprised by the results for the judgement on the media break: While less noticeable media breaks are expected when comparing pSV and EV, there should be

	all listeners			native listeners		
	pSV	CV	EV	pSV	CV	EV
media break	3.3	1.5	1.2	3.5	1.7	1.2
intelligibility	4.1	1.9	3.3	4.1	2.3	3.3
prosody	3.3	2.8	3.2	3.5	3.1	3.6
overall	3.4	2.1	3.4	3.4	2.4	3.6

Table 2: Mean opinion score results for pre-recorded prompts with standard voice (pSV), corporate voice (CV), and enriched corporate voice (EV).

no perceived break in CV, since this voice is created from a single speaker's data.

Whereas the enriched corporate voice EV outperforms CV in all categories and reaches the scores of pSV for both prosody and overall impression, there is a need for further improving the intelligibilty of EV. We are currently analyzing which parts of SV's database contribute most to the synthesis of variables used in our scenario, and plan to incorporate this data. However, in doing so we must take care in preserving the corporate speaker's voice characteristics in a larger common segment database.

5. Conclusions

In this paper we have addressed the creation of corporate synthetic voices from real IVR data. We discussed a couple of normalization steps that were found useful when employing a highly automated training procedure for the creation of the synthesizer's back-end speech database.

A second focus of this paper was on the creation of *enriched* corporate voices that pool data from different speakers in order to further improve the quality of output speech for text which is poorly represented in the corporate speaker's training data. In doing so, we could get independent of the availability of the corporate speaker and could avoid the recording of additional data. Initial listening test results show that by this approach we can compete with the quality of pre-recorded prompts, while at the same time simplifying the application.

Both a further improvement of synthesis quality and a broadening of the scope of the application seems possible through the incorporation of more data seems possible. This should come along with the use of state-of-the-art voice morphing techniques (see, for example, [13]) in order to preserve the corporate speaker's identity.

Acknowledgements. The authors would like to thank all participants in the listening tests, and would like to acknowledge the discussions with Ellen Eide and Raimo Bakis of IBM research, who also provided the spectral equalizer.

6. References

[1] A. Hunt and A. Black, "Unit Selection in a Concatenative Speech Synthesis System using a Large Speech Database", in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Atlanta, 1996, vol. 1, pp. 373–376.

[2] R. Donovan and E. Eide, "The IBM Trainable Speech Synthesis System", in *Proc. of the 5th Int. Conf. on Spoken Language Processing*, Sydney, 1998.

[3] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A.K. Syrdal, "The AT&T Next-Gen TTS System", in *Proc. of the Joint Meeting of ASA, EAA, and DAGA*, Berlin, Germany, 1999.

[4] Y. Kim, A. Syrdal, and M. Jilka, "Improving TTS by Higher Agreement Between Predicted Versus Observed Pronunciations", in *Proc. of the 5th ISCA Tutorial and Research Workshop on Speech Synthesis*, Pittsburgh, PA., 2004.

[5] M. Ostendorf and I. Bulyko, "The Impact of Speech Recognition on Speech Synthesis", in *Proc. of the IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, Ca., 2002.

[6] A. Black and K. Lenzo, "Limited domain synthesis", in *Proc. of the 6th Int. Conf. on Spoken Language Processing*, Beijing, 2000, pp. 411–414.

[7] V. Fischer, J. Botella Ordinas, and S. Kunzmann, "Domain adaptation methods in the IBM trainable speech synthesis system", in *Proc. of the 8th Int. Conf. on Spoken Language Processing*, Jeju Island, Korea, 2004.

[8] E. Eide, A. Aaron, R. Bakis, P. Cohen, R. Donovan, W. Hamza, T. Mathes, M. Picheny, M. Polkosky, M. Smith, and M. Viswanathan, "Recent Improvements to the IBM Trainable Speech Synthesis System", in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Hong Kong, 2003.

[9] M. Sakamoto and T. Saito, "An Automatic Pitch Marking Method Using Wavelet Transform", in *Proc. of the 6th Int. Conf. on Spoken Language Processing*, Beijing, 2000.

[10] W. Hamza, R. Bakis, and E. Eide, "Reconceiling Pronunication Differences between the Front-End and the Back-End in the IBM Speech Synthesis System", in *Proc. of the 8th Int. Conf. on Spoken Language Processing*, Jeju Island, Korea, 2004.

[11] W. Hamza and R. Donovan, "Data-driven Segment Preselection in the IBM Trainable Speech Synthesis System", in *Proc. of the 7th Int. Conf. on Spoken Language Processing*, Denver, 2002.

[12] E. Eide and M. Picheny, "Towards Pooled-Speaker Concatenative Text-to-Speech", in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Toulouse, 2006.

[13] H. Ye and S. Young, "High Quality Voice Morphing", in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Montreal, Canada, 2004.