ADAPTATION TECHNIQUES FOR SPEECH SYNTHESIS IN UNDER-RESOURCED LANGUAGES

Gopala Krishna Anumanchipalli, Alan W Black

Language Technologies Institute
Carnegie Mellon University, Pittsburgh, PA 15213
{gopalakr,awb}@cs.cmu.edu

ABSTRACT

This paper presents techniques for building speech synthesizers targeted at limited data scenarios - limited data from a target speaker; limited or no data in a target language. A resource sharing strategy within speakers and languages is presented giving promising directions for under-resourced languages. Our results show the importance of the amount of training data, the selection of languages and the mappings across languages in a multilingual setting. The objective evaluations conclusively prove that the presented adaptation techniques are well suited for building voices in resource-scarce conditions.

Index Terms: Speech Synthesis, Adaptation, Voice conversion, under-resourced languages.

1. INTRODUCTION

In today's digital age, there is an increasing use and acceptance of text-to-speech(TTS) technologies in the internet, mobile phones and dialogue systems. Besides, the use of speech as an output modality also enables information access for low-literate and visually impaired users. There is a compelling case for the development of speech synthesis technology in possibly all languages of the world. However, most languages have little or no resources required for building synthesis systems. Even for languages rich in speech and language resources, there is a need for efficient strategies for user-customization. Eliciting limited data ($<$ 2 mins) from the subject should sufficiently allow adaptation of an existing synthesizer to his voice. In this paper, we address both these situations as resource-scarce scenarios for bilding acceptable quality speech synthesizers.

While there is no definite notion of the minimum amount of resources required for training, availability of at least one hour of clean speech recordings is the norm for building high-quality functional speech synthesizers. This is in addition to phonetic and linguistic knowledge that requires annotated text resources in the language. This can be expensive and non-trivial for most languages. Many languages still have limited or no resources required to build text-to-speech systems. This makes building synthesis systems challenging using existing techniques. While, unit selection [10] continues to be the underlying technique in most commercial systems, its requirement of a large amount of well recorded and labeled speech data to ensure optimal unit coverage makes it prohibitive for under-resource situations. Statistical parametric synthesis [16], on the other hand is more liberal in its requirements, produces a more flexible voice comparable in quality to unit selection synthesis. Hence it is ideal for building voices in resource-scarce conditions.

Section 2 briefly describes our statistical parametric speech synthesis framework. A description of the resources required for building parametric voices follows in Section 3 including strategies for building voices under the resource-scarce conditions. Experiments and results are presented in Section 5.

2. STATISTICAL PARAMETRIC SPEECH SYNTHESIS

We use Clustergen [6], a statistical parametric framework within the Festvox [13] voice building suite. Fig. 1 shows a schematic representation of the training and testing phases in Clustergen. In the training phase, source and excitation parameters of the speech are extracted. Text-normalization and letter-to-sound(LTS) rules are applied on the transcription. The speech and phonetic transcriptions are automatically segmented using Hidden Markov Model (HMM) labeling. The speech features are then clustered using available phonetic and linguistic knowledge at a phoneme state level. Trees for duration, spectral (e.g. MFCC) and source (e.g. F0) features are built during the training phase. During testing (i.e. Text-to-Speech) input text is processed to form phonetic strings. These strings, along with the trained models are used to generate the feature parameters which are vocoded into a speech waveform by a synthesis filter (e.g. MLSA for MFCCs).

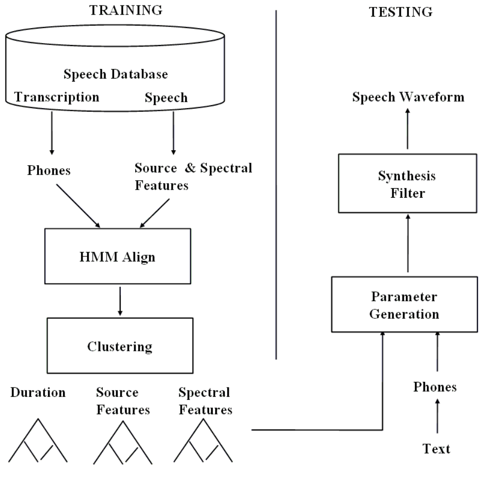

Fig. 1. Schematic diagram of the Clustergen framework

In this framework, models are stored as Classification And Regression Trees (CART) of the phone state. Each phone is realized as a left-to-right Markov chain of three states (roughly corresponding to the initial, middle and final states of a phone). The intermediate nodes of the tree are questions about phonetic and other high levels of contextual information (e.g., parts of speech). At the leaf nodes of the tree are the Gaussian codebooks corresponding to the feature instances falling in that path of the tree. The parametric representation (multi-dimensional Gaussians, in this case) makes transformations feasible via simple matrix algebraic operations. This flexibility of parametric models makes them well suited for adaptations required in under-resource conditions. Although this framework is similar to HTS [1], Clustergen generates the utterance frame by frame, rather than by state, allowing more detailed modeling.

3. BOOSTING RESOURCES FOR VOICE BUILDING

In this section, the resources required for building a voice are described, The specific alternatives for dealing with each kind of resource scarcity—that of limited target speaker data and target language data are presented in subsections 3.1 and 3.2 respectively. According to [11], the issues that need to be addressed while building a voice for a new language are 1) Definition of a phoneme set, 2) Creation of a lexicon and/or Letter-to-Sound(LTS) rules, 3) Text analysis, 4) Building prosodic models and 5) Building a waveform synthesizer.

For languages that have an IPA, SAMPA or a phoneset defined on another standard, they may be adequate to produce synthesizers of acceptable quality. However, for languages that have no established phonesets, it takes a few expert hours to design one based on the acoustic phonetic information of the phonemes in the language. For languages that are fairly phonetic (high grapheme-to-phoneme correspondence), grapheme-based phonesets have been shown to be adequate[17]. It should be noted that there is a certain arbitrariness in the allophonic variations within a language or even among speakers and there is no one best phoneset, optimal for all voices. Similarly, construction of a lexicon and LTS rules is non-trivial and the effort varies across languages, but a rule-based or a data-driven, statistical model for LTS has become commonplace for synthesizers in most languages [18]. In the following sections, the issues with limited amount of speech data are presented.

3.1. Limited data from a target speaker

As mentioned earlier, building a voice for a speaker requires a good amount clean recorded speech. It is thus desirable to have techniques that can work with just a few minutes of speech and produce good quality output. Recalling from Section 2, building a voice implies constructing decision trees for duration, source and spectral features. When the data is limited, phone coverage and contextual converge are both insufficient. This hurts any automatic technique to label the data. Even the estimated parameters (Gaussian means and variances) tend to be unreliable.

To compensate for this, data from one or more speakers may be used to build the 'source model' upon which the adaptation technique can impose the target speaker's identity.

This problem is studied extensively as 'model adaptation' proposed for speech recognition, starting with the work of [19], later also successfully applied for speech synthesis [20]. The selection of the source speakers on which to adapt may also be improved. Techniques involving speaker clustering and cohort selection have previously shown significant gains. There is also related work in voice transformation and feature space transforms [4] that deal with limited target speaker data.

3.2. Limited data in a target language

Lack of sufficient speech data for building speech systems is a common problem for most minority languages of the world. The GlobalPhone [8] project addresses this problem for speech recognition by exploiting existing resources in several languages to create a new language synthesizer. Similar attempts in speech synthesis [2] [14] also succeeded in creating a new language synthesizer sharing resources from several languages. This process is briefly described in the next section.

3.2.1. Multilingual Speech Synthesis

The 'source' voice in case of a target language adaptation is a multilingual voice. The training data for such a voice is speech included from several languages and the processed transcriptions in the respective languages. Since the phonetic properties (and labels) of the languages could be different, a global phoneset is created for the multilingual voice which assigns the same phonetic category to phonemes of different languages with the same acoustic phonetic properties. This strategy optimally shares the speech data across languages wherever appropriate. This also helps 'boost' the phonetic coverage of each language. However, this process requires carefully developed phone mappings between languages. The voice is built in a similar way as a monolingual voice after the mapping.

For the target language, the phoneset is mapped to that of the global set of the multilingual voice. The adaptation follows the same strategy as in a monolingual case transforming only the phonemes appropriate to the data presented for the target language. As shown in our results, the choice of the languages included in the training, and the amount of data in each language also affects the quality of the voice in a target language.

4. EVALUATION OF VOICES

We use Mel-Cepstral Distortion (MCD), a spectral distance measure proposed for evaluating voice conversion performance. It is given by the equation

$$MCD = 10/ln10\sqrt{2\sum_{d=1}^{24}(mc_d^{(t)} - mc_d^{(e)})^2} \qquad (1)$$

where $mc_d^{(t)}$ and $mc_d^{(e)}$ are the target and the estimated spectral vectors respectively. MCD is known to correlate well with the quality of a voice [12]. The significance of MCD is quantitatively shown as a function of the training data size. A reduction of 0.12 MCD is shown as being equivalent to doubling the amount of training data used for building the voice. This is shown to be consistent across speakers and languages. The MCD measure is hence relevant both in the limited target speaker and limited new language data in this work.

5. EXPERIMENTS AND RESULTS

In this section, we report our observations of the adaptation techniques in each limited data situation. In all experiments, 50 dimensional Mel-Frequency Cepstral Coefficients (static + delta features) are used as the spectral representation. The features are clustered using phonetic and contextual questions. For growing the CART trees thresholded with a stop value of 50 instances at the leaf node. All adaptations are done only on the spectral features. A simple z-score mapping is done for the fundamental frequency to adjust to the dynamic range of the target speaker.

5.1. Limited target speaker data

To evaluate the limited target speaker data scenario, we use varying amounts of adaptation data of an American male speaker taken from the arctic database [7]. As the source model, we use 41 American English speakers of the Wall Street Journal speech corpus [15]. An 'average' voice is built from 3 hours of speech data sampled evenly across 41 speakers. It is shown that such an average voice is closer to an arbitrary new speaker since it has the average characteristics of all training speakers, and tends to be speaker independent.

We report two experiments of voice adaptation, one model based, MLLR adaptation [19] and the other feature based using Joint density GMM-based estimation (GMM-JDE) [3]. Since the target data is limited, adaptation is done only on the Gaussian means and the original variances are retained.

Figure 2 shows the MCD of the estimated speech with respect to the reference data as a function of the amount of data used for adaptation. It can be seen that even with 20 utterances there is a significant improvement in the voice and it is closer to the target speaker. The two techniques begin almost giving same improvements, and begin to converge with increasing adaptation data. The GMM-JDE technique converges more quickly. MLLR outperforms the GMM-JDE technique when more adaptation data is presented. This shows that of the two techniques, MLLR exploits data more effectively for this task.

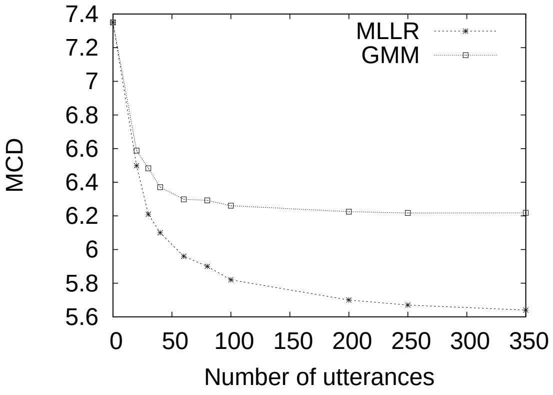

Fig. 2. Performance with increasing size of adaptation data from target speaker

5.2. Limited new language data

For simulating a limited new language data condition, a subset of the Globalphone database is selected. This subset consisted of 10 female speakers, one from each of Chinese (ZH), English (EN), German (DE), Japanese (JP), Croatian (KR), Portuguese (PT), Russian (RU), Spanish (ES), Swedish (SW) and Turkish (TR). Of these, German is set aside as a test target language. The remaining 9 languages are included in different amounts to also study the effect of data size in a multilingual setting. 10% of the sentences are set aside as testing data for each language.

Figure 3 presents the MCDs of the individual languages using the same multilingual voice. The x-axis is the amount of training data contributed by each language. The near-linear pattern of (es, pt, sw ru, en and zh) suggests that MCD (hence, voice quality) is proportional to the training data size, and this holds even in the multilingual setting. The good performance of Turkish and Japanese irrespective of the amount of training data may be explained by their simple phonetic structures.

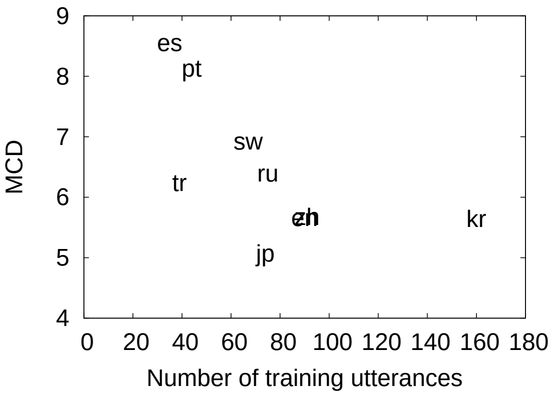

Fig. 3. MCDs of individual languages using a multilingual voice. Note: en/zh have the same amount of training data and the same MCD score.

For testing the new languages, we choose German (DE) and Telugu (TE) languages. The phonemes of these languages are mapped to their closest alternative from any of the nine different languages included as training. The overlap in the acoustic phonetic feature values of these phonemes are used to determine the closeness between phonemes (currently no weight is given to different acoustic phonetic features). The multilingual voice is incrementally adapted with data from the target language. Figure 4 shows the performance of the adaptation as MCD gains as a function of increasing amount of adaptation data. It can be seen that the German voice has a relatively lower MCD than the Telugu voice even without any adaptation. This may be explained by the fact that Telugu belongs to the Dravidian language family which is not represented in the training languages, while European languages are well represented. Informal listening tests also show that while the voices are understandable, they have new accents caused by the training languages.

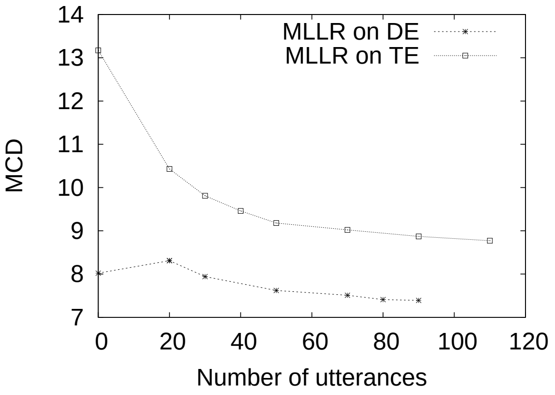

Fig. 4. German (DE) and Telugu (TE) language MCDs with increasing adaptation data

6. LANGUAGE SELECTION EXPERIMENTS

In this section, we report our experiments with changing the subset of languages included in training the multilingual voice. From, the initial subset of 9 languages chosen for training in the previous section, two subsets are created one including all but English and the other consisting of all but Chinese language. The choice of these languages is for two reasons: 1) They are phonetically quite distinct and 2) They contribute the same number of training sentences (as can be seen in the overlayed en/zh tags in the Fig. 3)

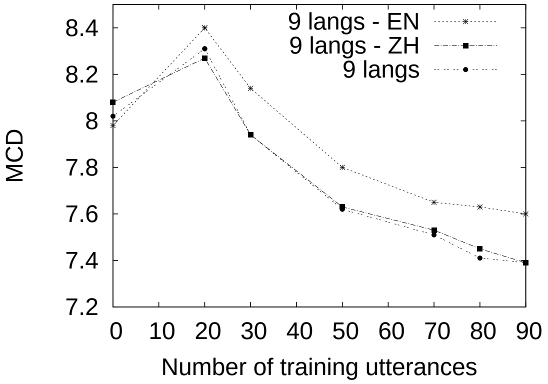

Fig. 5. German adaptation with different Training languages

From 6, as we expected the removal of English, a language phonetically similar to German, gives worse results, while the removal of Chinese, does not make much difference to the quality of German voice.

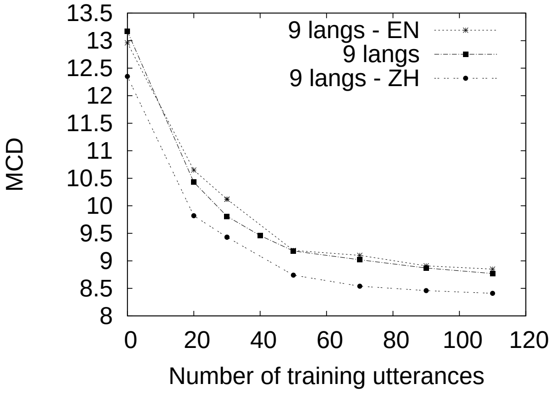

Fig. 6. Telugu adaptation with different Training languages

For the same experiment for creation of the Telugu voice, we find the removal of English does not make much difference, which we believe is due to the fact that Telugu and English are phonetically not very close. The unexpected result though is that the removal of Chinese improves the results. This shows that language selection is clearly important. One hypothesis for this result is the fact that Telugu has a larger number of stop distinctions than English (e.g. aspirated and unaspirated) such allophones do appear in English but are not phonetic. The initial models have these distinctions conflated, but become distinct with more adaptation data. However in Chinese, aspirated and unaspirated allophones do not occur within stops, hence the training data actually biases the initial phone models more and requires more training data to contract.

7. CONCLUSIONS

This work proposes adaptation techniques for under-resourced languages that clearly give promising results. The selection of initial models, although can be done by simple acoustic phonetic feature matching, our results show that more subtle selection of initial phonetic models and the languages that contribute to them may give even better results. We have yet to discover an efficient automatic method to improve these existing techniques.

The second important result is that the resulting synthesis quality seems to be linearly related to amount of training data, even across several languages.

8. ACKNOWLEDGMENT

This work was supported by the Fundação de Ciência e Tecnologia through the CMU/Portugal Program, a joint program between the Portuguese Government and Carnegie Mellon University.

9. REFERENCES

[1] K. Tokuda, T. Masuko, T. Kobayashi and S. Imai, “Speech Parameter generation from HMM using Dynamic Features”, *ICASSP ’95*, Detroit, USA, 1995.

[2] Javier Latorre, Koji Iwano and Sadaoki Furui, “New approach to the polyglot speech generation by means of an HMM-based speaker adaptable Synthesizer”, *Speech Communication*, 48:1227–1242, 2006.

[3] Tomoki Toda, Alan W Black and Keiichi Tokuda, “Voice Conversion based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory”, in *IEEE Transactions on Audio, Speech and Language Processing*, 15(8):2222-2236, 2007.

[4] Yannis Stylianou, O. Cappé and E. Moulines, “Statistical Methods for Voice Quality Transformations”, in *Eurospeech*, Madrid, Spain, 1995.

[5] Viet-Bac Le, Laurent Besacier, and Tanja Schultz, “Acoustic-Phonetic Unit Similarities for Context Dependent Acoustic Model Portability” in *ICASSP*, 2006, Toulouse, France.

[6] Alan W Black, “CLUSTERGEN: A Statistical Parametric Synthesizer using Trajectory Modeling”, *Interspeech* 2006, Pittsburgh, PA.

[7] John Kominek and Alan Black “CMU ARCTIC databases for speech synthesis”, Tech Report CMU-LTI-03-177, Carnegie Mellon Unversity `http://festvox.org/cmu_arctic`, 2003.

[8] Tanja Schultz “GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University”, *ICSLP* 2002, Denver, USA.

[9] Yanagisawa K. and Huckvale Mark, “A Phonetic Assessment of Cross-Langage Voice Conversion.”, *Interspeech* 2008, Brisbane, Australia.

[10] A. J. Hunt and Alan W Black, “Unit selection in a concatenative speech synthesis system using a large speech database”, *ICASSP* 1996, Atlanta, USA.

[11] Alan W Black and Kevin Lenzo, “Multilingual Text-to-Speech Synthesis”, *ICASSP* 2004, Montreal, Canada.

[12] John Kominek, Tanja Schultz and Alan W Black, “Synthesizer voice quality on new languages calibrated with mel-cepstral distorion”, *SLTU* 2009, Hanoi, Vietnam.

[13] Alan W Black and Kevin A. Lenzo, “Building Synthetic voices”, `http://festvox.org/`.

[14] Alan W Black and Tanja Schultz, “Speaker Clustering for Multilingual Synthesis”, *ITRW Multilingual Speech and Language Processing*, 2006, South Africa.

[15] D. Paul and J. Baker, “The design for the wall street journal based CSR corpus” *DARPA Speech and Natural Language Workshop*, 1992.

[16] Zen, H,. Black, Tokuda, K., and Black, A., “Statistical Parametric Speech Synthesis” *Speech Communication*, 51(11), pp 1039-1064, November 2009.

[17] Font Llitjos, A and Alan W Black “Unit Selection without a phoneme set” *IEEE TTS Workshop*, Santa Monica, USA, 2002.

[18] Maskey, S. , Tomokiyo, L. and Black, A. “Bootstrapping Phonetic Lexicons for New Languages” *ICSLP*, Jeju Island, Korea, 2004.

[19] Leggetter, C. J. and Woodland, P.C.. “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models” *Computer, Speech & Language*, pp. 171-185, 1995.

[20] M. Tamura, T. Masuko, K. Tokuda and T. Kobayashi, “Speaker adaptation for HMM-based speech synthesis system using MLLR,” *ESCA/COCOSDA Workshop on Speech Synthesis*, Nov, 1998.