Using Articulatory Features and Inferred Phonological Segments in Zero Resource Speech Processing

Pallavi Baljekar, Sunayana Sitaram,
Prasanna Kumar Muthukumar, and Alan W Black

Carnegie Mellon University, USA
pbaljeka@cs.cmu.edu, ssitaram@cs.cmu.edu, pmuthuku@cs.cmu.edu, awb@cs.cmu.edu

Abstract

Unsupervised discovery of subword units is an important problem in recognition and synthesis of zero-resource languages, in which phonesets may not be known and the only resource that may be available is speech. We use techniques that we have recently developed for building synthetic voices for very low resource languages without a written form to discover such units. We use Articulatory Features trained on labeled speech in a higher resource language to infer phonological segments of varying granularity. We use both the raw Articulatory Features and the Articulatory Features of the inferred units as frame-based representations of speech. We evaluate our techniques on minimal pair ABX discrimination within and across speakers. In addition, to exploit the duration information we get from the inferred phonological units, we also present evaluation results on Mel Cepstral Distortion, an objective metric of speech synthesis quality. We evaluate our techniques on multiple databases of English, and also on Tsonga and Indic languages, in which we apply the above methods cross-lingually.

Index Terms: unsupervised techniques, low resource, articulatory features

1. Introduction

Although speech processing has progressed significantly for languages with significant resources, there are still many languages for which well-defined phoneme sets, or even well-defined writing systems do not exist. Thus finding techniques that can give a useful, reliable, symbolic representation of recordings of human speech is still somewhat of an open task. Recent work has developed various frame-based acoustic representations that can be used to match different occurrences of instances of words and phrases, but in this paper we look at using higher level representations of the speech [1, 2].

This paper shows how previous work that we have done on developing an unsupervised symbolic representation of speech, suitable for "text" to speech systems for unwritten languages, may be suitable for recognition and matching tasks as well as just speech generation. Specifically we build on top of a phonetically derived acoustic representation of speech, [3], that we refer to as *Articulatory Features* (AFs). AFs, which can be derived from arbitrary streams of recorded speech, provide vectors of features values 0-1 that represent IPA-like phonetic features. Note our articulatory features might sometimes be called by others as phonetic features, and are not directly related to what we would call articulatory position features as might be discovered from an electro-magnetic articulograph. Our AFs are directly derived from speech in a language independent way, using standard software algorithms without any specialized hardware.

Finding a frame based AF representation is only the first part of our task. We then discover a segmental representation of the signal, that is phoneme-like derived from the AFs that is at least sufficient to reconstruct the signal using statistical parametric synthesis techniques.

Our initial work on text to speech for languages without a writing system, used *cross-lingual phonetic decoding* to come up with a phoneme-based written form for building TTS systems in languages without a standardized written form [4]. In our subsequent work we also used unsupervised and cross-lingual techniques to come up with higher level units [5], which improved objective and subjective measures of TTS system quality. However these techniques are still dependent on an originally seeded (cross-lingual) phonetic system. Our more recent work that we use here derives segment-based *Inferred Phones* (IPs) using acoustically derived frame-based AFs [6].

The rest of the paper is organized as follows. Section 2 relates this work to previous work. Section 3 describes the data and resources we used for our experiments. Section 4 contains details of the techniques we used, and describes the metrics used for evaluation, followed by results in section 5 and our inferences and conclusions in section 6.

2. Relation to prior work

The main goal of this work is to find the basic units in language in an unsupervised fashion. These smaller units could either be words, sub-word units, phonemes or even sub-phonetic units, which are maximally distinguishable in an ABX test [7][8].

Most methods in this domain have thus looked at variety of methods related to either unsupervised pattern discovery or unsupervised acoustic modeling . The first set of methods treat it as a pattern recognition problem, by first finding repetitive patterns in the database and then using these patterns to build word based models [9, 10, 11, 12] . In [2], they use a hashing scheme to convert the raw input features to a binarized fixed length form and then do a clustering of these fixed length vectors, with the main goal of improving feature based term discovery.

The second set of methods includes unsupervised acoustic modeling based approaches, wherein the speech is first segmented, then a clustering of these segments is carried out based on minimizing a certain objective measure and finally a retraining of the acoustic model is done. This process is repeated until convergence. In [13], the authors take a similar approach to sub-word modeling wherein, they train an auto-encoder to give encoder posteriors which are then binarized and clustered to a maximum of 64 units . These 64 units are then used to obtain a transcription of speech and based on this transcription the acoustic model is retrained and this process of segmentation, clustering and re-training continues until the model converges. In most systems these three sub-tasks of *segmentation*, *clustering* and *re-training* are carried out as independent tasks to one another.

However, the authors in [14] combine the segmentation step along with the clustering step by first starting out with a single HMM state to represent the entire dataset and then iteratively splitting these HMM states based on some objective measure. The authors in [15] go a step further, by jointly training an acoustic model using a nonparametric Bayesian model namely the Dirichlet process mixture model. However, most of these methods approach the task of unsupervised unit discovery from an ASR perspective, with the objective of increasing the classification or discrimination ability of each unit.

How our work is different is that we approach the problem from a synthesis viewpoint and so we are interested in finding the basic units in speech that are discriminable enough to be used to generate speech rather than classify between different phonetic or sub-word units. Thus the units our synthesis pipeline discovers are designed to be invertible and robust to speaker variation. We believe that this fits in nicely with the goal of the zero resource challenge where the aim is to mimic how a child learns language units in its infancy and is able to distinguish across speaker variability and retain the common units across speakers.

Our initial work on TTS without text relied on cross-lingual phonetic information. This method inherently makes assumptions about the phoneme distribution in the original cross-lingually trained phonetic models. Although relevant to the task we wanted to better represent the phonemes in the target (unlabelled) language.

Details of this technique are described in [6]. Our first stage was to build on the notion of AFs as described in [3]. Such features have been used beyond speech recognition in representing expressive speech [16] and cross-lingual voice conversion [17].

3. Data and resources

The data that we used for our experiments was provided by the organizers of the Zero Speech challenge was in English and Xitsonga. The English Buckeye database consists of 9 hours of data spoken by 12 speakers, with multiple speakers in the same audio file. Since the size of each audio file was many minutes long, we split the files into 10 second long files for some of our tasks and then recombined them during evaluation. We used the NCHLT Xitsonga speech corpus [18] provided by the organizers, which consists of 4.5 hours of speech by 24 speakers.

In addition, we used two other databases. The first was a combined database of the RMS and SLT Arctic data [19], which is around 2 hours of US English speech data, from one male and one female speaker. The second was a combined Hindi database with recordings from a local female speaker and the Blizzard challenge 2015 data [20], consisting of recordings of one male speaker, giving a total of around 2.5 hours of data.

For all our experiments, we used the (US English) WSJ acoustic model distributed with the CMU Sphinx toolkit [21] for cross-lingual phonetic decoding. We used a trigram German phonetic language model for decoding and performed multiple iterations of decoding and building targeted acoustic models from the decoded transcripts and the speech, as described in [4].

We built all our models in the context of the Festival Speech Synthesis Engine [22] and the Festvox voice building tools [23]. We built CLUSTERGEN [24] Statistical Parametric Synthesis voices so we could calculate the Mel cepstral distortion (MCD) [25], an objective measure of speech synthesis quality.

4. Experiments and Evaluation Methodology

In this section, we describe the details of the different models that we compared and how we evaluated these models.

4.1. Feature Description

4.1.1. Baselines

We used the MFCC features provided by the organizers, (which we refer to as the baseline Mceps) as well as the SPTK Mceps as baselines for this task. The baseline Mceps are 13 dimensional MFCC features computed every 10*ms* and the ABX score is computed using the cosine distance. The SPTK Mceps are derived using the SPTK toolkit [26] and are 50 dimensional vectors (25 dimensions + Δ) which are used in synthesis and hence designed to be invertible.

4.1.2. Z-model Mceps

The Z-model Mceps are speaker normalized Mceps. Each speaker's Mceps are mean and variance normalized to match the average across all speakers in the database.

4.1.3. Cross-lingual phonetic decoding

We decoded the speech from all the databases cross- lingually using the WSJ model and obtained phonetic transcripts. This process is done iteratively, with a targeted acoustic model being created at each iteration that is used to decode speech at the next iteration. Typically, we build voices at each iteration and measure the MCD of the voices. This iterative process is carried out till the MCD converges and stops improving. The iteration that produces the lowest MCD is selected as the best iteration. From our previous work we have found that the best labels are obtained in around iteration 3, so we chose the labels of iteration 3 for all our databases. The choice of labels is not critical here, since we only use the timestamps of these labels for the inferred phonemes.

4.1.4. Raw Articulatory Features

We trained a neural network on a large corpus of multi-speaker English speech [27]. This predicts a 26 coefficient vector of 0-1 values for phonetic features, such as voicing, nasality, place of articulation, etc, trained from the labeling derived by forced aligned models of the original WSJ data. This produces a frame-based labeling of Mel-cepstrum features.

4.1.5. Inferred Phonemes

Using these AF's, the next stage is to use a cross-lingual phonetic recognizer to discover similar segments (of varying length) in the acoustics. Then we take these segmentations and re-cluster them into similar segments based on their frame-level AFs as described in [6]. This is helpful because a cross-lingual phonetic recognizer may label all */k/* like sounds together, while, this post recognition re-clustering may separate out different types of */k/* (e.g. aspirated and unaspirated) into different segment-types. We can control the number of segment-types to find the number of symbols that can best re-construct the acoustic signal using statistical parametric text to speech techniques.

We refer to these segment-types as inferred phones.

4.2. Evaluation Metrics

The ABX metric measures the discriminative power of the subword units within and across speakers. For the across speaker measure in the ABX task, if we select an ABX triplet to be such that, A and B are triphones from the *same speaker*, having the same context, but varying in the middle phone, like *put* and *pat*, while X is the same as A except from a *different* speaker. The goal then is to find linguistic units, such that A and X are much closer than B and X. Similarly in the within speaker task, the goal remains the same, except that X is another instance of A from the *same* speaker.

In addition to using the ABX metric, we also used Mel Cepstral Distortion (MCD) of voices built with the representations that we inferred. To calculate the MCD, we hold out 10% of the data and build a synthetic voice using the rest of the data. Then, we resynthesize the held-out data and compare the Mceps to the Mceps of the original speech. The MCD is an objective metric commonly used to measure the quality of speech synthesizers and has been found to correlate with subjective metrics of synthesis quality.

Furthermore, we also did a word based comparison described in the next section, since our IPs did not fit into the ABX evaluation framework.

5. Results

First, we ran the ABX evaluation software provided by the organizers on the baseline Mceps, SPTK Mceps, Z-model normalized SPTK Mceps and the raw AFs that were extracted from the audio. Since the AFs were extracted frame-wise, we could directly use them with the evaluation software as they were. From Table 1 we see that the articulatory features perform much better than the Z-model Mceps across all databases. This indicates that the AFs are doing speaker normalization implicitly, and are more robust to speaker variation. We also see that the within-speaker error rate for the Z-models is slightly higher than the Mceps, which is expected, given that the Z-models are doing speaker normalization.

Table 1: *ABX on Mceps and AFs (% Error rate)*

Method	Data	Within Speaker	Across Speaker
Baseline Mceps	Buckeye	15.6	28.1
SPTK Mceps	Buckeye	16.69	29.50
Z-model Mceps	Buckeye	16.98	28.01
Raw AFs	Buckeye	18.35	29.84
Baseline Mceps	RMS+SLT	–	–
SPTK Mceps	RMS+SLT	7.56	18.45
Z-model Mceps	RMS+SLT	7.54	17.62
Raw AFs	RMS+SLT	7.53	15.02
Baseline Mceps	Tsonga	19.10	33.8
SPTK Mceps	Tsonga	19.69	33.93
Z-model Mceps	Tsonga	19.74	30.69
Raw AFs	Tsonga	18.12	29.73
Baseline Mceps	Hindi	–	–
SPTK Mceps	Hindi	8.89	28.13
Z-model Mceps	Hindi	9.11	27.22
Raw AFs	Hindi	8.33	24.91

Next, we used the raw AFs to create IPs as described earlier. Instead of using the raw AFs for each file as we had done before, we replace the IPs for a file with the vector of average value of the phoneme's AFs, calculated across the *entire* database. Since the ABX task was set up to be a frame based evaluation, we replicated this average value for each frame that the IP spanned.

Table 2 shows the ABX results on IPs of different sizes. The stop value was used to control the number of IPs that were inferred. Stop value of 1200, 1000 and 800 were experimented with. For the same stop value, the exact numbers of IPs varied across databases as can be seen in Table 2. As we see, none of the IPs were able to do better than the AFs or the Mcep baseline. We hypothesize that the major difference in the performance of the IPs on the Tsonga database as compared to its AF's is because of the nature of the Tsonga database which contains a higher variability of speakers, has been recorded over the telephone and consists of short utterances, a combination of all of which does not allow us the benefit of having clean data to train the cross-lingual model to give suitable IP representations. The AF's do perform better because, they firstly are frame based features and secondly, because they implicitly do speaker normalization.

Table 2: *ABX on IPs of different sizes (% Error rate)*

No. of IPs	Data	Within Speaker	Across Speaker
81	RMS+SLT	14.30	19.06
65	RMS+SLT	14.64	19.41
55	RMS+SLT	14.92	19.57
71	Tsonga	42.84	46.03
57	Tsonga	44.09	46.43
82	Hindi	20.83	29.62
55	Hindi	20.05	28.91

Since our work focuses in finding phoneme-like segments in untranscribed data, we would like to test these IPs within the ABX test framework used above. But that framework is not very appropriate for a sub-word segmental model. As it is tested against some phonetic-like truth, the segment size will be similar and thus our deduced segments will be about the same size (given some reasonable assumption about finding appropriate boundaries). Thus scores will be a simple 0 or 1, depending on whether it fits the frame exactly or not. Thus we also present some other measures that might better show our own contribution.

The main issue is measuring phoneme sized units against phoneme sized units when the boundaries are one of the key variances in such a model, thus it would be better to extend the size of the comparison units to something more like words (specifically multiple phoneme-like segments long).

We analyzed our data (for which we have true transcriptions) and looked for multi-syllable words that appear more than once. We then used these words as our test words. We then compare these words with each other within and across speakers using different measures. In the cases where we are comparing the same word the measure should be lower, and when they are different words the measure should be larger. We can do this with simple frame based parametrization (as done above) but as the words are longer we can also do this in the IP domain. Additionally we can also do this using synthesis, as we can generate an acoustic stream from the symbolic IP stream.

Table 3 column 1 lists the average DTW cost across all instances of the same speaker saying the same keyword, i.e., we are finding an average cost of matching the keyword in one sentence to all other instances of it and calculating an average of

Table 3: *Word-based scores-Within Speaker (DTW cost)*

Method	Data	Keyword	Not Keyword
Mceps	RMS-SLT	2.55 ± 1.89	4.44 ± 0.04
Average AFs	RMS-SLT	0.59 ± 0.55	0.92 ± 0.02
Synthesis	RMS-SLT	3.30 ± 5.05	3.67 ± 0.11
Mceps	Hindi	7.27 ± 9.11	8.78 ± 0.22
Average AFs	Hindi	0.93 ± 1.49	0.97 ± 0.05
Synthesis	Hindi	3.79 ± 3.95	3.84 ± 0.08

Table 4: *Word-based scores-Across Speaker (DTW cost)*

Method	Data	Keyword	Not Keyword
Mceps	RMS-SLT	2.22 ± 1.46	4.58 ± 0.04
Average AFs	RMS-SLT	0.41 ± 0.35	0.94 ± 0.02
Synthesis	RMS-SLT	1.59 ± 3.67	3.68 ± 0.11
Mceps	Hindi	10.08 ± 8.93	12.44 ± 0.30
Average AFs	Hindi	0.88 ± 1.53	1.10 ± 0.05
Synthesis	Hindi	4.45 ± 4.83	5.27 ± 0.14

this cost as compared to column 2 which represents the average of the cost when matched to other words apart from the keyword the same speaker said in the corpus. Table 4 column 1 compares the cost of measuring the keyword said by speaker 1 to all instances of the same keyword said by speaker 2 in the database, *vs.*, column 2 which lists the average cost of comparing keywords said by speaker 1 to all non keyword instances spoken by speaker 2. These measures within and across have been done as overall cost measures for Festival Mcep (baseline), Average AF's (the vector representation of the IP) and synthesized Mceps after rebuilding the voice from the unsupervised IP units obtained. We see that the IP as a feature for doing keyword spotting is successful, since in both cases across and within speaker it is able to give a lower cost on a simple DTW Euclidean distance metric. One interesting point to note from this result is that the variance is lower for AF's as compared to those of the Mceps and is again indicative of the speaker normalization that is happening implicitly in deriving this representation.

The motivation behind using the speech synthesis pipeline was to find a set of linguistic units which are good at representing the speaker agnostic, invertible sub-units in the speech corpus. The measure of how good these units are in generating speech can be measured with the MCD. Since the MCD is a distance based metric, lower is better, and it is database-specific, so it cannot be compared across the different databases.

Thus in addition to reporting the MCD scores for the voice built from our best IP, we have also reported scores from cross-lingual phonetic decoding from WSJ acoustic model. Since the MCD is database specific, we also give ground truth (Full TTS baseline) when transcripts were available for comparison. Table 5 lists the MCD of voices built with transcripts from cross-lingual phonetic decoding, our best IPs and the full knowledge-based speech synthesizer (Full TTS-groundtruth) – for comparison. An increase of 0.08 is found to be perceptually significant, while an increase of 0.12 is equivalent to doubling the data [28].

Here, we see that for English, the phonetic decoding MCD is better than the IP MCD. Although this may seem surprising, we must note that we used the WSJ acoustic model to decode the RMS-SLT voice, so this is not being done cross-lingually. So, the phones in the phonetic voice are appropriate for this voice, which results in a higher MCD. Both, the phonetic decoding MCD as well as the IP-based MCD are higher than the knowledge-based (full TTS) based MCD, which is to be expected. For Hindi, the IP-based voice has a lower MCD than the cross lingual phonetic voice which indicates that the IPs are a better representation of the speech for Hindi.

Table 5: *MCDs of voices built with different transcripts*

Data	Transcript	MCD
RMS-SLT	Full TTS	4.97
RMS-SLT	Phonetic Decoding	5.51
RMS-SLT	IPs	5.86
Hindi	Full TTS	4.94
Hindi	Phonetic Decoding	6.60
Hindi	IPs	5.94

6. Conclusion

In this paper we present an alternative unsupervised linguistic unit discovery method to find speaker agnostic, invertible speech units which are optimized for speech synthesis. We have investigated these proposed AF and IP based features as an alternative to unsupervised acoustic modeling and in the context of performing well on the ABX task.

However, since our proposed features do not fit well into the ABX framework, which requires the discovery of units which can fit within its framework of phoneme-sized ground truth, we have also reported MCD scores which measure how good the synthesis of the IP based voices is, which in turn measures the discriminability of the IP representation.

Although the IPs give a good symbolic representation of the speech they are still not the most ideal representation. As the number of segments in an utterance are initially derived from a cross-lingual phonetic recognizer, they most probably represent phoneme-sized units. It may be better to allow them to be split into multiple subsegments (the IP-based text to speech synthesizer automatically models sub-phonetic segments).

We find that on clean datasets, with less number of speakers, our proposed method works well. However, on noisy datasets like the Xitsonga dataset, which consists of many speakers and short utterances recorded via a telephone, we find that our model fails to perform as well , which we conjecture is due to the lack of good data to adapt the baseline model to.

The work presented here is still preliminary, a more elaborate speaker specific adaptation technique may help – though we have found that AFs are typically a better speaker independent representation. However, when synthesizing templates for matching, adapting the acoustics toward the target speaker in the utterance will improve performance.

Also IPs alone probably do not give all the information useful for word level matching. We know in IP-based text to speech that addition of word boundary information helps synthesis and thus finding super segmental information about syllable and word (like) boundaries will probably help higher level matching too (and certainly the generation of synthesized acoustics for later matching).

7. References

[1] K. Levin, K. Henry, A. Jansen, and K. Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings," in *IEEE Workshop on Automatic Speech*

Recognition and Understanding (ASRU). IEEE, 2013, pp. 410–415.

[2] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2011, pp. 401–406.

[3] F. Metze and A. Waibel, "A flexible stream architecture for ASR using articulatory features." in *INTERSPEECH*, 2002.

[4] S. Sitaram, S. Palkar, Y.-N. Chen, A. Parlikar, and A. W. Black, "Bootstrapping text-to-speech for speech processing in languages without an orthography," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7992–7996.

[5] S. Sitaram, G. K. Anumanchipalli, J. Chiu, A. Parlikar, and A. W. Black, "Text to speech in new languages without a standardized orthography," in *Proceedings of 8th Speech Synthesis Workshop, Barcelona*, 2013.

[6] P. K. Muthukumar and A. W. Black, "Automatic discovery of a phonetic inventory for unwritten languages for statistical speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2594–2598.

[7] T. Schatz, V. Peddinti, X.-N. Cao, F. Bach, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair ABX task (II): Resistance to noise," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[8] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline," in *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*, 2013, pp. 1–5.

[9] A. Jansen and K. Church, "Towards unsupervised training of speaker independent acoustic models." in *INTERSPEECH*, 2011, pp. 1693–1692.

[10] A. Jansen, S. Thomas, and H. Hermansky, "Weak top-down constraints for unsupervised acoustic model training." in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 8091–8095.

[11] G. Aimetti, L. ten Bosch, R. K. Moore, and N. Nijmegan, "The emergence of words: Modelling early language acquisition with a dynamic systems perspective," *Proceedings of EpiRob*, vol. 9, pp. 17–24, 2009.

[12] G. Aimetti, R. K. Moore, and L. ten Bosch, "Discovering an optimal set of minimally contrasting acoustic speech units: A point of focus for whole-word pattern matching," 2010.

[13] L. Badino, C. Canevari, L. Fadiga, and G. Metta, "An auto-encoder based approach to unsupervised learning of subword units," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7634–7638.

[14] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised learning of acoustic sub-word units," *ACL-08: HLT*, p. 165, 2008.

[15] C.-y. Lee and J. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 40–49.

[16] A. W. Black, H. T. Bunnell, Y. Dou, P. K. Muthukumar, F. Metze, D. Perry, T. Polzehl, K. Prahallad, S. Steidl, and C. Vaughn, "Articulatory features for expressive speech synthesis." in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4005–4008.

[17] B. Bollepalli, A. W. Black, and K. Prahallad, "Modelling a noisy-channel for voice conversion using articulatory features," in *INTERSPEECH*, 2012.

[18] N. J. De Vries, M. H. Davel, J. Badenhorst, W. D. Basson, F. De Wet, E. Barnard, and A. De Waal, "A smartphone-based ASR data collection tool for under-resourced languages," *Speech communication*, vol. 56, pp. 119–131, 2014.

[19] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[20] Blizzard Challenge. [Online]. Available: http://festvox.org/blizzard

[21] P. Placeway, S. Chen, M. Eskenazi, U. Jain, V. Parikh, B. Raj, M. Ravishankar, R. Rosenfeld, K. Seymore, M. Siegler *et al.*, "The 1996 Hub-4 Sphinx-3 system," in *Proc. DARPA Speech Recognition Workshop*, 1997, pp. 85–89.

[22] P. Taylor, A. W. Black, and R. Caley, "The architecture of the Festival speech synthesis system," in *In the Third ESCA Workshop in Speech Synthesis*, 1998, pp. 147–151.

[23] A. W. Black and K. A. Lenzo, "Building synthetic voices," *Language Technologies Institute, Carnegie Mellon University and Cepstral LLC*, 2003.

[24] A. W. Black, "Clustergen: A statistical parametric synthesizer using trajectory modeling." in *INTERSPEECH*, 2006.

[25] M. Mashimo, T. Toda, K. Shikano, and N. Campbell, "Evaluation of cross-language voice conversion based on GMM and STRAIGHT," in *The Seventh European Conference on Speech Communication and Technology (EUROSPEECH) Aalborg, Denmark*, 2001.

[26] S. Imai, T. Kobayashi, K. Tokuda, T. Masuko, K. Koishida, S. Sako, and H. Zen, "Speech signal processing toolkit (SPTK), version 3.3," 2009.

[27] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[28] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean Mel cepstral distortion." in *SLTU*, 2008, pp. 63–68.