Zero resource speech synthesis using transcripts derived from perceptual acoustic units

Karthik Panda D S, Hema A Murthy
Indian Institute of Technology Madras
{pandia,hema}@cse.iitm.ac.in

Abstract
Zero resource speech processing is the task of building vocabulary independent speech synthesis systems, where transcriptions are unavailable for training data. It is, therefore, necessary to convert training data into a sequence of fundamental acoustic units that can be used for synthesis during the test. This paper attempts to discover, and model perceptual acoustic units consisting of steady state, and transient regions in speech. The transients roughly correspond to CV, VC units, while the steady-state corresponds to sonorants and fricatives. The speech signal is first preprocessed by segmenting the same into CVC-like units using a short-term energy-like contour. These CVC segments are clustered using a connected components-based graph clustering technique. The clustered CVC segments are initialized such that the onset (CV) and decays (VC) correspond to transients, and the rhyme corresponds to steady-states. Following this initialization, the units are allowed to re-organize on the continuous speech into a final set of AUs in an HMM-GMM framework. AU sequences thus obtained are used to train synthesis models. The performance of the proposed approach is evaluated on the Zerospeech 2019 challenge database. Subjective and objective scores show that reasonably good quality synthesis with low bit rate encoding can be achieved using the proposed AUs.

Index Terms: acoustic unit discovery, speech synthesis, zerospeech

1. Introduction
Zero resource speech processing is a sub-field in speech processing which does not use any transcribed data for applications that generally mandate the availability of transcriptions to train models. Such applications include speech recognition, keyword spotting, document classification, text-to-speech synthesis, to name a few. Zero resource speech processing is useful for digital processing of languages that either have low audio resources, or languages that do not have a script. Additionally, AUs discovered can also give new insights into the nature of resources, or languages that do not have a script. Additionally, AUs discovered can also give new insights into the nature of resources, or languages that do not have a script. As the task is to synthesise speech using the acoustic units, perceptual units are superior to other kinds of units. Instead of just clustering the syllabic units as proposed by Räsänen and Nagarajan, we proposed AUs to model both the syllabic regions and inter-syllabic regions. The AUs that we model correspond to steady-state and transient regions in speech. To model the AUs, we first train the models on shorter syllable-like segments, which is motivated by children’s language acquisition. For this, the syllable-like units are used to initialize the models corresponding to onset, rhyme and coda. After the models are fine-tuned by training on syllable-like segments, the models are retrained on continuous speech to obtain final models.

The proposed approach is evaluated on zerospeech 2019 challenge [14] dataset. The results showed in this paper correspond to that of the systems submitted to the challenge. Zero resource speech processing coursework every two years focuses on zero-resource speech processing. The broad objective of the challenge is to construct an end-to-end spoken dialogue system for
an unknown language. The first two challenges [15, 16] focused on unit discovery and lexicon discovery. The third and the current challenge focuses on unit discovery and speech synthesis using the discovered units. Subjective evaluations show that zero resource speech synthesis is indeed possible and the results are comparable to a supervised system which uses transcribed audio for both acoustic modeling and synthesis.

The rest of the paper is organized as follows. Section 2 explains the proposed approach for acoustic unit discovery. Section 3 gives a brief overview of the experimental setup and data set used for experimentation. The results are shown and discussed in Section 4. Section 5 concludes the paper.

2. Proposed AUD technique

As the task is to synthesize speech using the discovered acoustic units, AUs of the form of perceptual units are intuitive than AUs of arbitrary form. Units of the size of syllables are accepted to be the basic units of speech perception. Unlike phonemes and context-dependent phones, approximate syllabic units can be obtained by processing the signal without the need for the availability of transcription. The peaks and valleys in the envelope of the speech signal naturally segment the speech signal into syllable-like units. Instead of directly clustering such syllabic segments to obtain syllable-like AU, as in case of Nagarajan and Räätänen, the inter-syllabic regions are explicitly modeled.

The overall flow of the proposed unit discovery technique is shown in Figure 3. The speech is first segmented into syllable-like units using a processed short-time energy function. The syllable-like segments are clustered by applying dynamic time warping (DTW) between all possible segment pairs. K-nearest neighbour connected component clustering approach [17] is applied to cluster similar segments. For each segment, the k-nearest segments are identified by using the DTW scores. A graph is then constructed with segments as nodes. An edge exists between two nodes i and j if i is in the set of the k-nearest neighbours of j and vice-versa. All the segments in a cluster are homogeneous. A snippet of three clusters along with the units is shown in Figure 1. An arrow between a node A to another node B is present if B is in the k-nearest neighbour of A. The segments in a cluster is sonorant regions and hence can spawn more than two words. The units belonging to each cluster are contained in the corresponding shaded regions. The figure shows that the units within a cluster are highly homogeneous.

The segments are assumed to contain three distinct units characterized by a vowel onset, rhyme and a vowel offset as illustrated in Figure 2. The segments in a cluster correspond to a sequence of three symbols consisting of an onset, rhyme and offset. The assumed units are used as transcripts for all the segments in a cluster, and Hidden Markov model (HMM) is used to train the individual units. The models are realigned and retrained in a self-training fashion until the overall likelihood of the data with
respect to the model converges. Let $O$ correspond to feature vectors and $W_{\text{true}}$ corresponds to the true label sequence, and $\Theta$ represents HMM parameters. Both $\Theta$ and $W_{\text{true}}$ are unknown quantities. The initial label sequence for the CVC units are assigned from the respective clusters. Given the label sequence, the model parameters, $\Theta_{\text{new}}$ are inferred as,

$$
\Theta_{\text{new}} = \arg \max_\Theta P(O, W_{\text{old}} | \Theta)
$$

A new label sequence is obtained, given the updated $\Theta_{\text{new}}$.

$$
W_{\text{new}} = \arg \max_W P(O, W | \Theta_{\text{new}})
$$

This self-training approach is similar to that of [3]. As the duration of the segments is short, this approach is useful to obtain a good initial model although inter-segment transients are not initialized. The obtained AU models are then used to transcribe the continuous utterances of long duration. This will help to train transients in the inter-segment regions. Similar to the first stage, a self-training approach is applied to train models. The obtained model is used in system 1. The number of units obtained by this approach is much larger than the actual perceptual units in any language. Hence, it should be possible to obtain a smaller set of AU. The AUs are merged using k-mean clustering. The obtained model is used in system 2. The number of units in the proposed approach is not fixed before unit discovery. The acoustic units depend upon the number of clusters obtained by the initial clustering. For system 1, the number of units obtained are 170 and 112 for development and test datasets respectively. For system 2, the units are again clustered to 40 units. This number is tuned on development dataset by running for different values. Mel frequency cepstral coefficients extracted with a frame size of 25ms and frame shift of 10 ms are used as feature vectors. Different systems are built with static, delta cepstral features with mean subtraction, LDA with MLLT, FMLLR-based speaker adaptation. Kaldi toolkit [20] is used for feature extraction and training models. To train synthesizer models, Merlin [21], an ANN based speech synthesis toolkit is used. As subjective evaluation is a time-consuming process, each team was allowed to submit a maximum of two systems. To shortlist the two systems, the multi-stimulus test with hidden reference and anchors (MUSHRA) [22] test was used.

Previous challenges used minimal pair ABX discrimination [23, 24] score to evaluate the quality of the discovered units. As this Zerospeech 2019 is on speech synthesis, apart from ABX score, the primary subjective evaluation metrics include intelligibility in terms of character error rate (CER), speaker similarity, and overall quality of the synthesis in terms of mean opinion score (MOS). CER (0-1) is obtained by manually transcribing the synthesized audio and comparing the transcripts with the reference transcription. Speaker similarity (0-5) shows the similarity of the synthesized voice to that of the target speaker.

### 4. Results and discussion

The baseline approach is similar to that of [3]. As the duration of the segments is short, this approach is useful to obtain a good initial model although inter-segment transients are not initialized. The obtained AU models are then used to transcribe the continuous utterances of long duration. This will help to train transients in the inter-segment regions. Similar to the first stage, a self-training approach is applied to train models. The obtained model is used in system 1. The number of units obtained by this approach is much larger than the actual perceptual units in any language. Hence, it should be possible to obtain a smaller set of AU. The AUs are merged using k-mean clustering. The obtained model is used in system 2. The number of units in the proposed approach is not fixed before unit discovery. The acoustic units depend upon the number of clusters obtained by the initial clustering. For system 1, the number of units obtained are 170 and 112 for development and test datasets respectively. For system 2, the units are again clustered to 40 units. This number is tuned on development dataset by running for different values. Mel frequency cepstral coefficients extracted with a frame size of 25ms and frame shift of 10 ms are used as feature vectors. Different systems are built with static, delta cepstral features with mean subtraction, LDA with MLLT, FMLLR-based speaker adaptation. Kaldi toolkit [20] is used for feature extraction and training models. To train synthesizer models, Merlin [21], an ANN based speech synthesis toolkit is used. As subjective evaluation is a time-consuming process, each team was allowed to submit a maximum of two systems. To shortlist the two systems, the multi-stimulus test with hidden reference and anchors (MUSHRA) [22] test was used.

Previous challenges used minimal pair ABX discrimination [23, 24] score to evaluate the quality of the discovered units. As this Zerospeech 2019 is on speech synthesis, apart from ABX score, the primary subjective evaluation metrics include intelligibility in terms of character error rate (CER), speaker similarity, and overall quality of the synthesis in terms of mean opinion score (MOS). CER (0-1) is obtained by manually transcribing the synthesized audio and comparing the transcripts with the reference transcription. Speaker similarity (0-5) shows the similarity of the synthesized voice to that of the target speaker.

### 3. Dataset and experiments

The proposed approach for speech unit discovery followed by speech synthesis is evaluated on the database provided as part of the zerospeech 2019 challenge. Parameter tuning is performed on development data, and the same procedure is repeated on the test data. The language for the development data is English, and the Indonesian language is used as test data. More details about the test corpus can be found at [18, 19]. Table 1 summarizes the dataset used for the experiments. Train unit dataset was provided to perform AUD and acoustic modelling. The duration of train data for both development and test languages are approximately 15 hours with 100 speakers in development data and 112 speakers in test data. The synthesis models are trained on a target voice. For the development data, one male and one female voices were provided, whereas for the test data, one female voice was provided. The audio provided in the test data has to be synthesized in the target speaker’s voice. The speakers in the source data are disjoint from that of the train unit data.

Unlike the baseline approach, the number of units in the proposed approach is not fixed before unit discovery. The acoustic units depend upon the number of clusters obtained by the initial clustering. For system 1, the number of units obtained are 170 and 112 for development and test datasets respectively. For system 2, the units are again clustered to 40 units. This number is tuned on development dataset by running for different values. Mel frequency cepstral coefficients extracted with a frame size of 25ms and frame shift of 10 ms are used as feature vectors. Different systems are built with static, delta cepstral features with mean subtraction, LDA with MLLT, FMLLR-based speaker adaptation. Kaldi toolkit [20] is used for feature extraction and training models. To train synthesizer models, Merlin [21], an ANN based speech synthesis toolkit is used. As subjective evaluation is a time-consuming process, each team was allowed to submit a maximum of two systems. To shortlist the two systems, the multi-stimulus test with hidden reference and anchors (MUSHRA) [22] test was used.

Previous challenges used minimal pair ABX discrimination [23, 24] score to evaluate the quality of the discovered units. As this Zerospeech 2019 is on speech synthesis, apart from ABX score, the primary subjective evaluation metrics include intelligibility in terms of character error rate (CER), speaker similarity, and overall quality of the synthesis in terms of mean opinion score (MOS). CER (0-1) is obtained by manually transcribing the synthesized audio and comparing the transcripts with the reference transcription. Speaker similarity (0-5) shows the similarity of the synthesized voice to that of the target speaker.

### Table 1: Dataset used for experiments

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Development language (English)</th>
<th>Surprise/Test language (Indonesian)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># Speakers</td>
<td># Utterances</td>
</tr>
<tr>
<td>Train unit</td>
<td>100</td>
<td>5941</td>
</tr>
<tr>
<td>Train voice</td>
<td>1 male</td>
<td>970</td>
</tr>
<tr>
<td>Test</td>
<td>24</td>
<td>455</td>
</tr>
</tbody>
</table>
the acoustic units to phonemes. In the figure, the blue regions correspond to transient and green regions correspond to steady-state. The figure shows that the AU transcription is better than phoneme transcription obtained from supervised training using TIMIT dataset. For instance, for the speech segment he spoke, the supervised model wrongly recognized the vowel /ow/ for the vowel /ow/. Even though the transients across different speakers have identical acoustic units across speakers, the symbols of the vowel-steady state regions are not identical. This can mean that, in the absence of supervised training, the same vowel belonging to different speakers with high variability in fundamental frequency may not be modelled as a single unit. Even the supervised system transcribed the vowels in an utterance uttered by two different speakers as different. It is a well known fact that the recognition accuracy of the consonants is very low for any phoneme recognizer. As the proposed approach models the consonants with context exclusively, the consonantal transients (CV and VC) are recognized with high recognition accuracy, as observed in Figure 4.

AUD problem can also be compared to that of children’s language acquisition [26]. During language acquisition, the syllable acts as the basic units that a child learns by continuous exposure. After the basic syllabic units are learnt, a child starts learning complex patterns that make up continuous speech. The proposed approach uses a similar technique for AUD, wherein first the training is confined to syllable-like units. Subsequently, the models are used to bootstrap continuous speech training.

Modelling of speech as transient and steady-state acoustic units is in accordance with other linguistic studies [27] that show that the steady-state vowels and CV, VC transients are basic units of speech perception. Massaro, in his studies, has classified the CV transients as stop transient, nasal transient, fricative transient. The clustered units in system 2 indeed aggregates similar kind of transients into one cluster. Even though for the synthesis task, the results from the subjective evaluation is better for system 1, it is observed that for other task such as spoken term detection, the system 2 outperforms system 1.

5. Conclusions

In the absence of transcribed data to train acoustic models for speech recognition and speech synthesis, the syllabic structure present in speech is valuable information. This information can not only be used as initial segments but can also be used to discover units which are perceptual. The results show that such units, when modelled using simple GMM-HMM framework, can achieve good synthesis quality and speaker similarity with reduced bit-rate.
6. References


