Discovering steady-state and transient regions in speech

Karthik Pandia D S
Indian Institute of Technology Madras
pandia@cse.iitm.ac.in

1. Problem statement

Acoustic modelling is the most crucial step in automatic speech recognition. Acoustic unit modelling techniques can be supervised or unsupervised. Supervised approaches suffer from the following issues: 1) require accurate transcriptions to train models, 2) the methods are usually language dependent and sensitive to the train data, and 3) rely upon large vocabulary along with its pronunciation dictionary. Hence an unsupervised approach is preferred when adequate data is not available to train the models.

Audio-query based keyword search problem is a classic case where unsupervised acoustic unit modelling techniques can play an important role. One or more audio snippets are used to search for the occurrences in an audio search file. Un-supervised audio-query-based keyword search either works by directly matching spectral feature vectors or by using the models trained using a set of discovered acoustic units (AU). In a supervised training scenario, context-dependent (CD) phones are considered to be reliable acoustic units. CD phones are longer units (≈ 300ms) typically consisting of three phonemes. A supervised approach uses text to align the audio with the CD phones. The problem is to discover reliable acoustic units which are similar to CD phones.

We propose a novel approach to discover AUs based on the transient and steady-state regions present in a speech signal. These units are V, CV and VC units, motivated by psycholinguistic studies. Vowels correspond to steady-state units, and CV and VC are transient units. To discover these AUs, we use syllable-like (CV∗VC∗) segments to initialize the models, followed by an iterative clustering and re-segmentation. To check the effectiveness of the discovered units, we use them on audio-query based keyword search or spoken term detection problem.

2. Motivation

As mentioned before discovering CD-phone-like units is preferred for unsupervised audio-query-based search task. The speech units V, CV, and VC have the desired properties hence can be potential AUs. The idea of using the steady-state and transient regions are also perceptually motivated. The V, CV, and VC are indivisible units of speech perception [1]. Further, the phonemes that correspond to consonants are not perceptually functional unless it is placed in a vowel context. Vowel only sentences with consonants replaced by noise are perceptually intelligible in word/sentence recognition [2–5]. But this is not true in the case of perception of isolated words; In the absence of context (language model), the consonants are equally important for perception. This is also in agreement with the working of ASR. A remarkable word recognition performance can be achieved using an ASR in spite of poor phoneme recognition rate. Such success is attributed to the language model or the context in the speech as in the case of the perceptual experiments. Hence, for unsupervised speech recognition, it is essential to recognize the consonants in addition to vowels reliably. This can be achieved by using the transients associated with the consonants.

Figure 1: Signal and vowel posterior for a TIMIT sentence

3. Steady-state and transient acoustic units

There are several attempts to discover acoustic units (AU) directly from speech [6–16]. Most approaches claim to discover phoneme-like AUs but in most cases, the units discovered loosely correspond to CD phones. The analyses of the discovered acoustic units show that the units are mostly vowels in different acoustic context [7, 11, 16] or syllable-like. The prime difference between the supervised techniques and unsupervised techniques is that supervision aids the segmentation in the former methods. To be precise, during the training, 1. the acoustic segments are constrained within the audio, 2. the sequence of phonetic identity is known. The segments are iteratively refined over the entire training corpus.

The existing AUD techniques are either top-down [6, 7] approaches, bottom-up [8–10] approach or cluster the pre-segmented audio iteratively [11, 13, 15–18]. The first two approaches do not pre-segment the audio allowing too much flexibility during the training. Even though the latter approaches segment the audio before clustering, they do not use the phonetic sequence to aid the training.

Table 1: Steady-state and transient acoustic units

<table>
<thead>
<tr>
<th>Segment Type</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transient</td>
<td>Pause</td>
</tr>
<tr>
<td>Transient</td>
<td>Vicar</td>
</tr>
<tr>
<td>Transient</td>
<td>Approximant</td>
</tr>
<tr>
<td>Silence</td>
<td>-4</td>
</tr>
</tbody>
</table>

The proposed approach tries to make use of the two advantages that a supervised ASR has. As mentioned earlier, we propose to discover steady-state and transient units. This is achieved by constraining the units within syllable-like segments. Since the definition of a syllable is \( CV^*VC^* \), it can be split as CV, V, and VC. From the perspective of the acoustic signal, a syllable is composed of onset, attack, and decay (OAD). As vowels can be considered to be stationary in a speech signal, there is always a transition before and after the stationary
A sequence matching is performed by transcribing both the query and search audio in terms of the proposed AUs. The proposed AUD technique is evaluated on query based audio search (spoken term detection) problem using the TIMIT dataset as shown in Figure 2. The first five entries in the table indicate the results of different AUD techniques using by applying S-DTW [8] on the posteriors of the AUs. Sequence matching is performed using the phoneme sequence obtained from BUT [19] English phoneme recognizer. The last row shows the results of sequence matching using the AU tokens of the proposed approach. Bayesian unsupervised approach [11], which is superior to all other methods, uses a supervised pre-segmenter to eliminate the unlikely boundaries. In spite of sequence matching using tokens, the proposed approach achieved 0.51 P@N, which is close to most frame-based approaches.

4. Experiments and discussions

The proposed acoustic unit discovery technique is applied on the TIMIT database. The discovered units are time aligned with that of the TIMIT phonemes to analyze the nature of the units. A close look at the discovered acoustic units reveals that the acoustic units indeed correspond to the steady-state and transients regions in speech. Massaro [1] have classified the consonant transients into three broad categories: stop, nasal and fricative transients. It is interesting to see in Figure 1 that the AUs correspond to these transients along with approximant transients. The individual cluster can be further classified into linguistic sub-categories like VC and CV clusters in transients and front and back vowels in case of steady-states.

AUs corresponding to 3 instances of the keyword money and one instance each for the keywords problem and children are shown in Figure 2. The IDs inside the segment correspond to the AUs and the IDs in the bottom correspond to the TIMIT phone transcription. The first row in the figure shows three instances of the keyword money. The obtained steady-state units c21, (c28, c23), and c11 correspond to the phone /m/, /a/ and /i/, and the transient units c17, c30 and c31 correspond to the transitions /m/ - /ah/, /ah/ - /a/ and /a/ - /iy/ respectively. The cluster c30 corresponds to nasal transients ( /ah/-/n/ and /bcl/-/em/) units which can be seen in the third instance of money, problem and children.

5. Future plans and road-map for the thesis

- [1] suggests that the place of articulation can be identified as the articulators rapidly travel (transition) from the consonant target positions to or from the contiguous steady-state vowel. As the proposed acoustic units have these characteristics, they can be used to extract articulatory features and can be used for speech processing.
- The segmentation can also be extended to attention-based models for speech recognition. The current attention based sequence-to-sequence models [20] use fixed size window for attention, and there is no single window size which has been identified to be appropriate. Using CVC structure to define the attention window is viable and also is intuitive.
6. References


