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. Unsupervised 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 ($\sim$ 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 (C^*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.

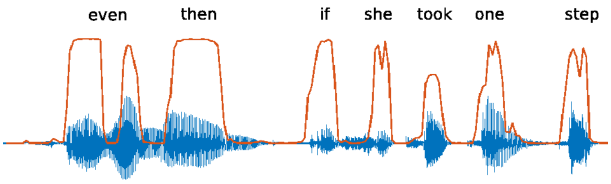

Figure 1: *Signal and vowel posterior for a TIMIT sentence*

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*

AU category	Type	#clusers	Examples
Steady-state	Vowel	9	ao ay ah ao aa ey eh ae ix iy ih ux ow
Transient	Stops	3	bcl bcl_b dcl dcl_d gcl gcl_g kcl kcl_k kcl_t pcl pcl_p tcl tcl_t
Transient	Nasal	3	aa_n ae_m ae_n ah_m ah_n ao_n ax_m ax_n ay_m eh_n en ey_m ey_n ih_n ih_ng ix_n ix_ng iy_n m m_eh m_ey m_ih m_ix n n_aa n_ae n_dcl n_dh ng
Transient	Fircative	9	ah_z axr_s axr_z ax_s ay_z ch dcl_jh eh_s er_z ey_s ey_sh h#_sh ih_s ih_sh ih_z ix_s ix_sh ix_z iy_s iy_sh iy_z jh m_z ow_z s sh tcl tcl_ch tcl_s ux_s ux_z z
Transient	Approximant	4	aa_l ao_l ax_l axr axr_w ax_w eh_l er ih_l iy_w l l_ae l_ah l_ax l_ay l_ey l_ih l_ix l_iy ow_l p_l r r_ey r_ih r_ix r_iy w w_ah w_ao w_ay w_eh w_ey w_ix w_iy
Pause	-	2	h# pau h#_h#
Mix	-	10	-

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 C^*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

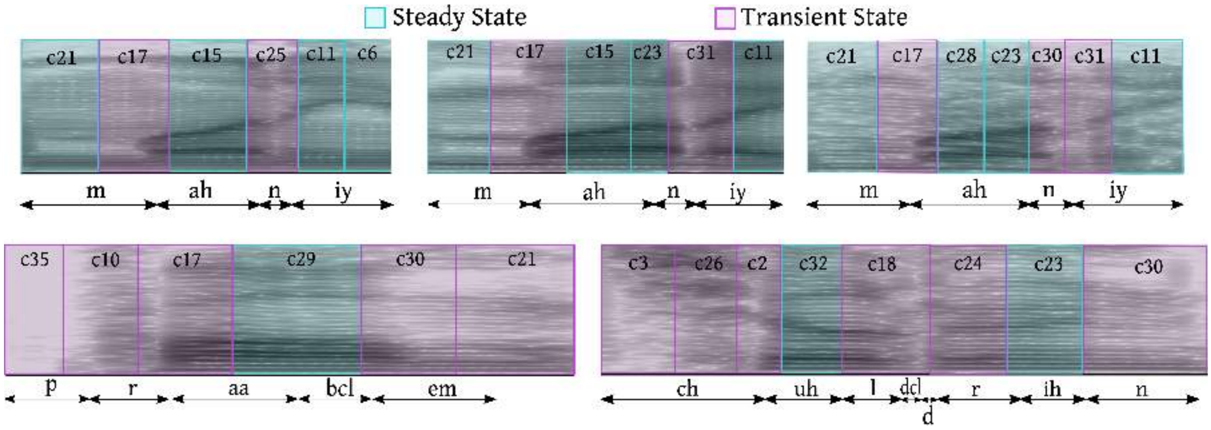

Figure 2: *Transient and steady-state segmentation of different keywords: money, problem and children*

region. These transitions are the onset and decay. At a gross level, the speech signal can be considered as being made up of rising, falling transients and stationary regions.

A vowel posterior function is obtained from a language-independent phoneme recognizer as shown in Figure 1. The segments thus obtained consist of a vowel in the center and non-vowel on both sides. Similar segments are grouped by applying dynamic time warping between all pairs. The grouping process is a bottom-up hierarchical agglomerative clustering. Each group contains a homogeneous set of CVC-like segments corresponding to three acoustic units labelled as OAD. Given the homogeneous CVC segments and their corresponding label sequence, the three assumed units are modelled using continuous density hidden Markov models (CD-HMM). These trained initial models are refined using a much larger collection of segments by aligning and retraining iteratively. This iterative modelling technique has the advantage as that of supervised CD phone training by assuming the presence of acoustic sequence as well as by constraining the training to CVC-like segments.

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 the phoneme /m/, /a/ and /iy/, and the transient units c17, c30 and c31 correspond to the transitions /m/-/ah/, /ah/-/n/ and /n/-/iy/ respectively. The cluster c30 corresponds to nasal transients (/ah/-/n/ and /bcl/-/em/) units which can be seen in the the third instance of *money*, *problem* and *children*.

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 devised in [8]. Precision at top N retrieval P@N is used as a keyword retrieval evaluation measure. The result of the proposed approach is compared with other AUD techniques in the literature as shown in Table 2. The first five entries in the table indicate the results of different AUD techniques using by applying SDTW [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.

Table 2: *Average P@N on TIMIT corpus*

Feature	Approach	P@N
GMM [8]	S-DTW	0.53
DBM [9]		0.51
Bayesian unsupervised [11]		0.63
DHDPHMM [13]		0.56
HDPHMM [13]		0.61
Phoneme tokens (BUT)	sequence matching	0.39
Proposed AU tokens		0.51

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 consonantal 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

[1] D. W. Massaro, "3 - acoustic features in speech perception," in *Understanding Language*, D. W. Massaro, Ed. Academic Press, 1975, pp. 77 – 124.

[2] R. A. Cole, Y. Yan, B. Mak, M. Fanty, and T. Bailey, "The contribution of consonants versus vowels to word recognition in fluent speech," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2, May 1996, pp. 853–856 vol. 2.

[3] F. Chen, L. L. N. Wong, and E. Y. W. Wong, "Assessing the perceptual contributions of vowels and consonants to mandarin sentence intelligibility," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. EL178–EL184, 2013.

[4] D. Fogerty and L. E. Humes, "Perceptual contributions to monosyllabic word intelligibility: Segmental, lexical, and noise replacement factors," *The Journal of the Acoustical Society of America*, vol. 128, no. 5, pp. 3114–3125, 2010.

[5] D. Kewley-Port, T. Z. Burkle, and J. H. Lee, "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 122, no. 4, pp. 2365–2375, 2007.

[6] A. Jansen *et al.*, "Weak top-down constraints for unsupervised acoustic model training," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 8091–8095.

[7] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised learning of acoustic sub-word units," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, ser. HLT-Short '08. Association for Computational Linguistics, 2008, pp. 165–168.

[8] Y. Zhang *et al.*, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *IEEE Workshop on Automatic Speech Recognition Understanding*, Nov 2009, pp. 398–403.

[9] Y. Zhang, R. Salakhutdinov, H. A. Chang, and J. Glass, "Resource configurable spoken query detection using deep boltzmann machines," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 5161–5164.

[10] H. Wang, C. C. Leung, T. Lee, B. Ma, and H. Li, "An acoustic segment modeling approach to query-by-example spoken term detection," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 5157–5160.

[11] C.-y. Lee *et al.*, "A nonparametric bayesian approach to acoustic model discovery," in *50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ser. ACL '12, 2012, pp. 40–49.

[12] H. Wang *et al.*, "Acoustic segment modeling with spectral clustering methods," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 264–277, Feb 2015.

[13] A. H. H. N. Torbati and J. Picone, "A nonparametric bayesian approach for spoken term detection by example query," *CoRR*, vol. abs/1606.05967, 2016.

[14] H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li, "A graph-based gaussian component clustering approach to unsupervised acoustic modeling," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[15] O. Räsänen, G. Doyle, and M. C. Frank, "Unsupervised word discovery from speech using automatic segmentation into syllable-like units," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[16] C. Chung, C. Chan, and L. Lee, "Unsupervised discovery of linguistic structure including two-level acoustic patterns using three cascaded stages of iterative optimization," *CoRR*, vol. abs/1509.02208, 2015.

[17] M. Huijbregts, M. McLaren, and D. van Leeuwen, "Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 4436–4439.

[18] G. Lakshmi Sarada, A. Lakshmi, H. A. Murthy, and T. Nagarajan, "Automatic transcription of continuous speech into syllable-like units for indian languages," *Sadhana*, vol. 34, no. 2, pp. 221–233, Apr 2009.

[19] P. Schwarz *et al.*, "Towards lower error rates in phoneme recognition," in *Text, Speech and Dialogue*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 465–472.

[20] R. Prabhavalkar, T. N. Sainath, B. Li, K. Rao, and N. Jaitly, "An analysis of attention in sequence-to-sequence models," in *Proc. Interspeech 2017*, 2017, pp. 3702–3706. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-232