PHONOLOGICAL FEATURE BASED VARIABLE FRAME RATE SCHEME FOR IMPROVED SPEECH RECOGNITION

Abhijeet Sangwan and John H.L. Hansen

Center for Robust Speech Systems (CRSS)
Department of Electrical Engineering
The University of Texas at Dallas, Richardson, Texas, U.S.A

ABSTRACT

In this paper, we propose a new scheme for variable frame rate (VFR) feature processing based on high level segmentation (HLS) of speech into broad phone classes. Traditional fixed-rate processing is not capable of accurately reflecting the dynamics of continuous speech. On the other hand, the proposed VFR scheme adapts the temporal representation of the speech signal by tying the framing strategy with the detected phone class sequence. The phone classes are detected and segmented by using appropriately trained phonological features (PFs). In this manner, the proposed scheme is capable of tracking the evolution of speech due to the underlying phonetic content, and exploiting the non-uniform information flow-rate of speech by using a variable framing strategy. The new VFR scheme is applied to automatic speech recognition of TIMIT and NTIMIT corpora, where it is compared to a traditional fixed window-size/frame-rate scheme. Our experiments yield encouraging results with relative reductions of 24% and 8% in WER (word error rate) for TIMIT and NTIMIT tasks, respectively.

1. INTRODUCTION

Speech production is a highly non-uniform phenomena in terms of the information flow rate. While both vowels and stops qualify as phonemes, they exhibit highly disparate temporal and spectral structures [1]. Hence, it is quite natural to assume that speech recognition would benefit by moving away from the standard inflexible spectro-temporal representation provided by MFCCs (mel frequency cepstral coefficients) and the popular framing scheme of 25ms window size/10ms skip rate.

While there is a ready consensus on the need for a more flexible feature representation scheme, the actual means of achieving this is a subject of research. In this paper, we develop a new technique towards variable framing rate (VFR) which is both consistent and meaningful by design. Our framing strategy is based on a discriminative treatment of broad phone classes, where we attempt to match a frame size and skip rate to the expected phonetic information flow rate. In other words, we perform high level segmentation (HLS) of speech into very broad phone classes, and tie a particular framing strategy with each class. The segmentation is assisted by extracting and training a set of relevant phonological features (PFs) [2].

The main argument for supporting VFR is the under representation of certain phonetic/linguistic content under the conventional framing policy. In general, relatively rapid acoustic events such as obstruents, especially if they are components of functional words (e.g., *th*e i*f* *h*im etc.) in a sentence are more ill-represented owing to their relatively informal articulation effort and short duration. Evidently, the under-representation of such phones form too few frames of acoustic evidence, which is not sufficient to warrant successful recognition from the ASR engine. On the hand, the ASR engine often confuses these frames as being part of the succeeding or preceeding words in an utterance, and thereby causes a mixture of deletion as well as substitution errors. A straight-forward method of mitigating under-representation is by employing a smaller frame size and rate around these islands of ambiguity. While a smaller frame size and skip rate does not necessarily result in better spectral representation, it certainly increases the number of frames used in capturing the information while simultaneously being more adept at tracking the evolution of the speech signal. This amounts to more meaningful acoustic evidence pointing towards the actual phonetic content.

The organization of this paper is as follows: In Sec. 2, we review the popular VFR strategies followed by our proposed method in Sec. 3.We describe our experimental setup in Sec. 4, and discuss the speech recognition results in Sec. 5.

2. REVIEW OF VFR SCHEMES

In past research, methods employed towards VFR are either driven by prior knowledge of phone class association of speech frames, or by information theoretic (IT) measures such as entropy. Traditionally, the former strategy is a back-end approach which is a direct attempt at solving the problem, since it attempts to customize the framing strategy to the phonetic

This study was supported by RADC under grant A40104.

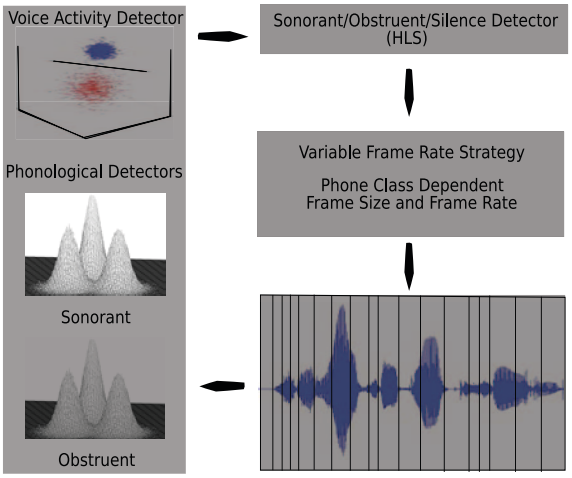

Fig. 1. Variable Frame Rate Scheme.

content of the speech frame. In [1], the authors developed an N-best list re-scoring strategy based on VFR where framing was guided by the phonetic output hypothesis of the ASR engine itself. The output hypothesis of the ASR engine was used to form 3, 5, 7 and 11 broad phone classes, and a series of increasingly refined framing schemes were adopted for each division resulting in multiple acoustic models. The final output hypothesis of the entire system was a weighted combination of all the individual frame rate acoustic models.

Other popular VFR strategies employ a number of front-end measures of difference (or similarity) such as short term entropy [3], frame by frame Euclidean distance [4, 5], and energy variations [6]. All of these measures are tuned towards segregating transients and steady state components of speech. In these methods, multiple thresholds are pre-assigned, and usually an N-ary detection is performed to guide the framing process. The major advantage of these techniques is that an explicit segmentation of the speech signal is not required. On the other hand, these techniques require heuristic tuning of the thresholds which might reduce their general applicability. Furthermore, if it is believed that VFR works due to its ability in dealing with large temporal variations in phonetics, then the best HLS of speech must also establish the theoretical upper limit on the performance of the IT driven approaches. Under this assumption, HLS seems to be a more direct methodology of solving VFR, if it is available prior to framing.

In this paper, we use prior knowledge of HLS to guide our framing process, and implement the VFR within the speech recognition front-end itself. In comparison to back-end techniques, this approach allows for a reduced complexity and faster system.

3. PROPOSED VFR SCHEME

As shown in Fig. 1, the proposed VFR scheme first performs a 3-ary HLS of speech into sonorants, obstruents, and silence,

Algorithm 1 Proposed HLS scheme

Obtain the VAD decisions V_d.
Obtain $\Lambda_N(\mathbf{z} \in G_i)$ using (2).
if $V_d == 0$ **then** ▷ (To obtain raw decisions)
  $D_r = 0$ ▷ (silence frame)
else if $V_d == 1$ **then**
  if $\Lambda_N(\mathbf{z} \in G_1) > \Lambda_N(\mathbf{z} \in G_2)$ **then**
    $D_r = 1$ ▷ (sonorant frame)
  else if $\Lambda_N(\mathbf{z} \in G_1) <= \Lambda_N(\mathbf{z} \in G_2)$ **then**
    if $\Lambda_N(\mathbf{z} \in G_2) <= 0.33$ **then**
      $D_r = 0$ ▷ (silence frame)
    else
      $D_r = 2$ ▷ (obstruent frame)
    end if
  end if
end if
Form group decisions D_i^G by grouping successive similar raw decisions D_r.
All obstruents, D_2^G flanked by pause, D_0^G on both sides are forced to pause.
loop for every D_i^G, $i = 0, 1$ ▷ (Tag unreliable D_r)
  if duration of $D_1^G <= 50ms$ **then**
    D_1^G is unreliable
  end if
  if duration of $D_0^G <= 90ms$ **then**
    D_0^G is unreliable
  end if
end loop
loop for every unreliable D_i^G ▷ (Correct unreliable D_r)
  if neighbors of D_i^G belong to same group **then**
    Assign D_i^G to same group as neighbors
  else
    Assign D_i^G to obstruents.
  end if
end loop

and then uses a preassigned phone-class based framing strategy. The HLS of speech is accomplished by means of a voice activity detector (VAD) and PFs (phonological features). The VAD scheme is based on competitive Neyman-Pearson (CNP) hypothesis testing which was earlier presented in [7]. The VAD output is denoted as V_d, and the values $V_d = 0$ and $V_d = 1$ represent pause and speech, respectively. The pause decisions of the VAD are retained and speech decisions are further processed to obtain sonorant/obstruent classification. The PFs used in the proposed VFR scheme are trained to distinguish between sonorants and obstruents. In contemporary literature, a number of classification techniques such as artifical neural networks (ANNs), GMMs (Gaussian Mixtures Models), HMMs (Hidden Markov Models) and Dynamic Bayesian Networks (DBNs) [8–10] have been employed to train PF detectors. In this paper, we employ a set of PFs based on 256-mixture gender-independent GMMs to distinguish between sono-

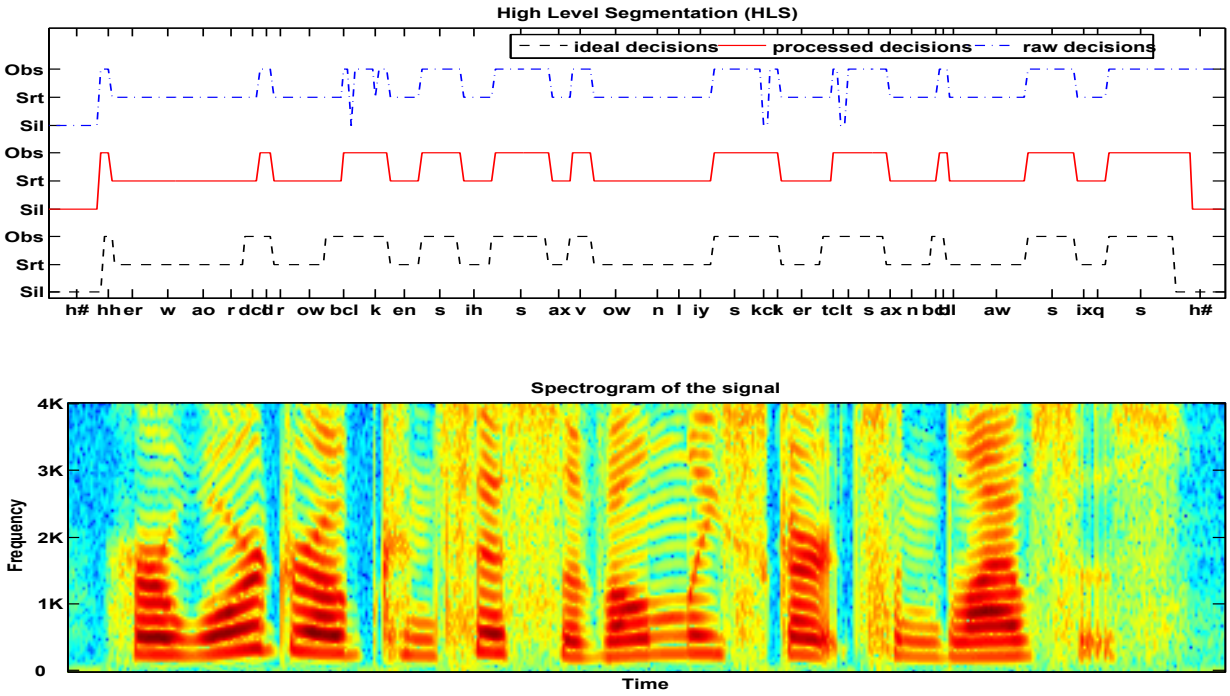

Fig. 2. Illustrating the three-way detection between sonorant, obstruents and silence for a TIMIT sentence: "Her wardrobe consists of only skirts and blouses." Ideal decisions represent the phone level alignments available for TIMIT.

rants and obstruents. Furthermore, unlike most PF based systems which employ MFCCs as the standard spectral representation, we experimented with alternate feature representations to choose the most robust scheme. In our experiments, we found that the PMVDR (perceptual minimum variance distortionless response) based PF scheme gives the best performance in terms of sonorant/obstruent classification [11].

The procedure for the HLS scheme is described below. We denote the GMM models for sonorant, obstruent and silence to be G_1, G_2 and G_3, respectively. Next, the likelihood of a speech frame $\mathbf{z}$ belonging to GMM G_i is given by:

$$\Lambda(\mathbf{z} \in G_i) = p(\mathbf{z}|G_i), i = 1, 2, 3. \quad (1)$$

In order to make useful comparisons across different speech frames, we normalize the likelihood scores, (i.e., we define

$$\Lambda_N(\mathbf{z} \in G_i) = \frac{p(\mathbf{z}|G_i)}{\sum_{j=1}^{3} p(\mathbf{z}|G_j)}, i = 1, 2, \quad (2)$$

as the normalized scores for sonorant (G_1) and obstruent (G_2) models). The speech frames are labeled as sonorants or obstruents by performing a simple comparison between the normalized likelihoods ($\Lambda_N(G_1)$ and $\Lambda_N(G_2)$) and choosing the maximum value as the correct phone class. The 3-ary decisions obtained at this point are termed as the raw decisions (D_r). The raw decisions for a TIMIT sentence are shown in Fig. 2. The ideal decisions for the same sentence built from the TIMIT transcriptions is also included for comparison. It can be seen that while the raw decisions show good sonorant detection, they contain numerous obstruent/silence confusions. Hence, further processing of raw decisions becomes necessary in order to improve the overall detection rate. To achieve this, we first reassign obstruent frames with normalized likelihood values below 0.33 to silence. This enables the HLS scheme to recover some silence frames which were mistaken for obstruents by the VAD and PFs. Now, similar contiguous decisions are further aggregated into blocks of continued sonorant, obstruent or silence periods. We denote these grouped decisions or periods of silence, obstruent, and sonorant as D_0^G, D_1^G, and D_2^G, respectively. Aggregating the similar decisions allows us to exploit durational constraints on speech production in correcting obvious errors in frame level decisions. First, we recover silence frames that are erroneously tagged as obstruents by finding all D_2^G (obstruents) that are flanked by D_0^G (silence) on both sides. Since the occurance of isolated observations of obstruents is highly unlikely, these frames must actually belong to silence or non-speech sounds such as a lip-smack, throat clearing etc. Next, all D_i^G are scanned and their component frames are tagged as reliable or unreliable decisions. Herein, all obstruent groups are assigned as reliable; and sonorant groups that last 50ms or less, as well as silence groups that last 90ms or less are tagged as unreliable. Again, sonorants and silence periods lasting for such short durations are unlikely and therefore believed to be erroneous decisions. Finally, the unreliable decisions are corrected in the following manner. The frames of an unreliably tagged group are

assigned to the neighoring reliable group if both neighbors belong to the same phone class. On the other hand, if the two neighboring decision groups belong to different phone classes, then the current unreliable group in question is always assigned to obstruents. The design of the above-described HLS scheme largely helps in resolving the ambiguity between silence and obstruents. In Fig. 2, the processed HLS final decisions are also shown for the same TIMIT sentence. Finally, all the steps of the HLS algorithm are summarized in the pseudo-code shown as Algorithm 1.

Upon obtaining the final decisions in terms of the assigned phone class, the speech signal is framed using a pre-assigned VFR strategy. Finally, in order to smooth the transitions into and out of obstruents, the detected transitions are always advanced and delayed by two frames, respectively. This also permits the ASR to capture transient information which is believed to be critical towards obstruent perception. It may be useful to note that the proposed VFR scheme forms a generic framework within which more refined phone-class segmentation, and a variety of framing schemes are possible.

4. EXPERIMENTAL SETUP

In order to evaluate the VFR scheme, we established a continuous speech recognition task for the TIMIT and NTIMIT corpora using the Sphinx speech recognition engine. The TIMIT corpus was suitably downsampled to 8kHz prior to training or testing. For both corpora, we use content dependent phone models with 600 senonically tied-states and diagonal covariance matrices [12]. Furthermore, the HMM topology used for modeling was a 5-state left-to-right model with no state skipping. In each experiment, the utterances are preemphasized with a factor of 0.97 and then appropriately framed. Subsequently, each frame is Hamming windowed and 13 dimensional MFCCs (mel frequency cepstral coefficients) are extracted for each frame. For obtaining MFCCs, the mel-scale is simulated using a set of 40 triangular filters. Furthermore, the delta, and delta-delta of the MFCC were concatenated to the static vector to form a single 39-dimensional feature vector. Cepstral mean substraction (CMS), variance normalization and automatic gain control (AGC) were also employed as part of the overall system. For the purpose of our experiments, we also employed a trigram language model.

In order to test the effectiveness of VFR across different speech features, we also employ PMVDR and WDCTC (warped discrete cosine transform cepstrum [13]) as alternate speech features to MFCC in our experiments. Since the VFR focuses on refining the temporal representation of speech, we expect the VFR to be beneficial across all spectral representations. Using the above-mentioned speech recognition system and speech features, a baseline performance of 12.1%, 12.7% and 10.2% WER (word error rate) is obtained for TIMIT using MFCC, PMVDR and WDCTC, respectively. Similarly, a baseline of 16.7%, 16.2% and 17.4% WER is obtained for NTIMIT using MFCC, PMVDR and WDCTC.

Table 1. Raw and Processed Detection Accuracy on TIMIT

TIMIT	Raw Detected			Processed Detected		
INPUT	Sil	Snt	Obs	Sil	Snt	Obs
Sil	**39.22**	4.9	56.05	**83.24**	3.55	13.25
Snt	0.35	**96.34**	3.32	0.77	**94.95**	4.29
Obs	9.60	21.32	**68.98**	13.02	19.71	**67.17**
Avg	**68%**			**81%**		

Table 2. Raw and Processed Detection Accuracy on NTIMIT

NTIMIT	Raw			Processed Detection		
	Sil	Snt	Obs	Sil	Snt	Obs
Sil	**23.79**	17.77	58.61	**74.52**	10.36	15.17
Snt	0.36	**94.39**	5.26	1.72	**92.66**	5.64
Obs	5.30	35.96	**58.62**	19.55	33.02	**47.27**
Avg	**59%**			**71%**		

(Sil: Silence, Snt: Sonorant, Obs: Obstruent)

In our experimental setup, we found that distinguishing between obstruents on one side, and silence/sonorants on the other gives the greatest benefit to the ASR performance. It seems that the ASR system is quite effective at distinguishing between sonorants and silence, and requires assistance in separating obstruents and silence. Hence, for our experiments we choose a standard frame rate of 25ms window/10ms skip rate for sonorants and silence, and a shorter 10ms/5ms for obstruents. It is worth mentioning that the need to distinguish between sonorants and silence may become more relevant in speech with more higher noise backgrounds than that observed in TIMIT and NTIMIT.

5. RESULTS AND DISCUSSIONS

The frame level detection accuracy of the sonorant and obstruent PFs are illustrated in Table 1 and 2 for TIMIT and NTIMIT, respectively. The benefit in processing raw PF decisions is clear, with processed decisions showing a relative improvement of 40% and 30% detection accuracy over the baseline performance. It is observed that the HLS scheme is able to recover numerous silence-obstruent confusions as intended by its design. The most important shifts in performance from Table 3 and 4 represent the reduction of silence detected as obstruent (56% to 13%) for TIMIT and (58% to 15%) for NTIMIT. Since silence and sonorant are treated equivalently by the frame settings (size and skip rate), confusions among them do not impact performance. It is also useful to note that since the HLS scheme always advances and delays all non-obstruent obstruent and obstruent non-obstruent boundaries, respectively by two frames, the VFR scheme is inherently tolerant towards errors in segmentation.

Recognition results for TIMIT and NTIMIT using the pro-

Table 3. TIMIT: VFR compared to conventional framing

TIMIT		Acc	Sub	Del	Ins	WER
MFCC	Fixed Rate	86.2	8.8	3.3	1.7	**12.1**
	VFR	89.6	6.7	2.5	1.1	**9.2**
PMVDR	Fixed Rate	85.6	9.3	3.3	1.7	**12.7**
	VFR	88.9	7.3	2.5	1.3	**9.7**
WDCTC	Fixed Rate	89.1	6.0	4.2	0.7	**10.2**
	VFR	91	5.5	2.7	0.8	**8.2**

Table 4. NTIMIT: VFR compared to conventional framing

NTIMIT		Acc	Sub	Del	Ins	WER
MFCC	Fixed Rate	81.4	12.2	4.5	1.9	**16.7**
	VFR	82.7	11.3	4.4	1.6	**15.7**
PMVDR	Fixed Rate	81.9	11.9	4.3	1.8	**16.2**
	VFR	82.8	11.6	4.2	1.5	**15.7**
WDCTC	Fixed Rate	81.6	11.2	6.2	1.0	**17.4**
	VFR	82.9	10.9	5.2	1.1	**16.0**

(Acc: Word Recognition Accuracy, Sub: Substitution, Del: Deletion, Ins: Insertion, WER: Word Error Rate)

posed VFR and conventional framing are shown in Tables 3 and 4, respectively. Using the proposed VFR scheme, we obtain a relavant WER improvement of 24%, 23.6% and 20% for MFCC, PMVDR, and WDCTC on TIMIT. Similarly, we obtain relative WER improvements of 6%, 3% and 8% on NTIMIT. Furthermore, it is observed that the VFR scheme results in an improvement in all error types (, i.e., substitutions, deletions and insertions), except for WDCTC where the insertions increase slightly. However, the increase in insertions for WDCTC is nominal owing to very low baseline insertion rate. On the other hand, we observe the greatest relative reductions in deletion errors followed by substitutions as a result of using VFR for TIMIT across all features. In the case of NTIMIT, the reductions in substitution errors is more significant than deletions. The results therefore show the clear benefit of the proposed variable frame rate method using phonological features for high level segmentation based ASR.

6. CONCLUSION

In this paper, a new variable frame rate (VFR) scheme was proposed as an alternative to the conventional fixed-rate framing. The proposed VFR scheme is based on high level segmentation (HLS) of speech where appropriately trained (phonological features) PFs were employed to segment speech into broad phone classes. The entire VFR scheme was implemented as the front-end of our speech recognition engine, which resulted in an efficient and cost-effective system. The proposed scheme was evaluated for speech recognition of the TIMIT and NTIMIT corpora, and benchmarked against the conventional fixed frame rate scheme. The proposed VFR scheme demonstrated very encouraging relative WER improvements of 24% and 8% on the TIMIT and NTIMIT tasks.

7. REFERENCES

[1] V. R. Gadde, K. Sonmez, and H. Franco, "Multirate ASR models for phone-class dependent N-best list rescoring," *ASRU*, Jun 2005.

[2] J. Frankel and S. King, "Articulatory speech recognition," *Interspeech*, Aug 2001.

[3] H. You, Q. Zhu, and A. Alwan, "Entropy-based variable frame rate analysis of speech signals and its application to ASR," *ICASSP*, Jan 2004.

[4] P. Le Cerf and D. Van Comperolle, "A new variable frame rate analysis method for speech recognition," *IEEE Signal Processing Letters*, Jun 1994.

[5] J. de Veth, L. Mauuary, B. Noe, F. de Wet, J. Sienel, L. Boves, and D. Jouvet, "Feature vector selection to improve ASR robustness in noisy conditions," *Eurospeech*, Jun 2001.

[6] J. Epps and E. H.C. Choi, "An energy search approach to variable frame rate front-end processing for robust ASR," *Interspeech*, Jul 2005.

[7] A. Sangwan, W.-P.Zhu, and M.O. Ahmad, "Design and performance analysis of Bayesian, Neyman-Pearson and competitive Neyman-Pearson voice activity detectors," *IEEE Trans on Sig. Proc.*, accepted for publication.

[8] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech and Language*, May 2007.

[9] F. Metze and A. Waibel, "A flexible stream architecture for ASR using articulatory features," *Interspeech*, Jun 2002.

[10] J. Frankel, M. Weser, and S. King, "Articulatory feature recognition using dynamic bayesian networks," *Interspeech*, Jun 2004.

[11] U. Yapanel and J.H.L. Hansen, "A new perspective on feature extraction for robust in-vehicle speech recognition," in *Eurospeech'03*, 2003.

[12] J. M. Huerta, "Speech recognition in mobile environments," *Phd Dissertation, CMU*, April 2000.

[13] R. Muralishankar, A. Sangwan, and D. O'Shaughnessy, "Statistical properties of the Warped Discrete Cosine Transform Cepstrum compared with the MFCC," in *EUROSPEECH'05, Lisbon, Portugal*, Sept. 2005.