Phoneme Dependent Frame Selection Preference

Tingyao Wu, Jacques Duchateau, Dirk Van Compernolle

Dept. ESAT, Katholieke Universiteit Leuven
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{Tingyao.Wu, Jacques.Duchateau, Dirk.VanCompernolle}@esat.kuleuven.be

10.21437/Interspeech.2007-10

Abstract

In previous study we proposed algorithms to select representative frames from a segment for phoneme likelihood evaluation. In this paper we show that this frame selection behavior is phoneme dependent. We observe that some phonemes benefit from frame selection while others do not, and that this separation matches the phonetic categories. For those phonemes sensitive to frame selection, we find that selecting frames at some pre-defined positions in the segment enhances the discrimination between phonemes. These phoneme-dependent positions are explicitly retrieved and used in a phoneme classification task. Experimental results on the TIMIT phonetic database show that the frame selection method significantly outperforms decoding by the classical Viterbi decoder.
Index Terms: frame selection, phoneme classification

1. Introduction

Most speech processing algorithms decompose speech signals frame by frame at a fixed frame rate, and each frame is represented by a fixed identical set of features such as mel-cepstra. Local frame based distance metrics or probabilities are computed and then integrated over time to yield a global score to evaluate the matching between test utterance and reference model. Although the frame-by-frame scheme is widely used probably due to the simplicity, researchers commonly agree that it is not an optimal solution. Many studies have proven that different frames play different roles in the recognition phase. These studies include variable frame rate [1], duration normalization [2], maximum likelihood based frame selection [3, 4], smoothed energy based frame selection [5], etc.

In our former study [4], we have shown that deliberately selecting only one frame to represent one state of a Hidden Markov Model (HMM) is sufficient to achieve pretty good performance. We also showed that at a low model complexity, a phoneme model trained by selected frames is more discriminative than the model using all frames. At least two factors influence the precision of spotting frames, namely time position and likelihood probability. While the likelihood probability is a measurement to assess the distance to different hypotheses, the position of a frame usually reveals different discriminative information: for example, in [6], the authors claim that frames at phoneme boundaries carry more speaker-related information, and those in steady state are more affiliated to speech issues. In fact, we observe that the frame position plays a more important role than the frame likelihood in frame selection. That is probably because the position factor reflects the global spectral trajectory while the likelihood factor is sometimes influenced by local optimal peaks.

An underlying problem in our former study is that all phonemes are forcefully represented by their individual three-frame sets, given the three-state HMMs. Motivated by the study, we further intuitively conjecture that not all phonemes are suitable for frame selection. For instance, discarding a few frames at boundaries of a vowel segment may decrease the noisy influence from context; but for a plosive, due to great variation within a segment, neglecting any part probably leads to a damage of global spectral reconstruction. Moreover, for the frame selection sensitive phonemes, the individual optimal number of selected frames and their positions may also differ. In this paper, we will empirically investigate the preference for frame selection for all phonemes by statistical tools, and explore phoneme-specific position sets for the frame selection sensitive phonemes. We will show that the classification behavior under frame selection fits the known phonetic categories rather well. The phoneme-dependent position sets are employed in a phoneme classification task and the experimental result shows that the discrimination between phonemes is enhanced.

This paper is organized as follows. In section 2 we will describe the database and the pre-processing used in this study. The approach for checking the sensitivity of a phoneme for frame selection is explained and assessed in section 3. In section 4.2, the result of the phoneme classification is presented and finally the conclusions are drawn.

2. Database and pre-processing

The phonetic TIMIT database is used to check the sensitivity for frame selection and to retrieve the phoneme-dependent position sets. All "sa" sentences are excluded from the training and recognition because they skew the phoneme occurrences, leaving 3696 sentences for training and 1344 for test.

We adopt the same phoneme set and the same evaluation method as described in [7]. Yet there is an essential difference for dealing with unvoiced/voiced closures: in [7] the closure was isolated from the following plosive, and classified into the silence category in the recognition phase; but in our settings, closure is folded into the next plosive as it behaves phonetically as a distinguishing part of the plosive.

A series of twelve dimensional mel-cepstra plus the energy, extracted from 30ms-length speech frames with 10ms frame rate, along with their velocity coefficients and acceleration coefficients is used for generating context-independent phoneme HMMs. Each phoneme is modeled by a three-state left-to-right context-independent HMM, with 16 gaussians per state. Phoneme boundaries are obtained from the available manual segmentation.

This work is sponsored by the FWO project G.0260.07 TELEX: Combining acoustic TEmplates and LEXical modeling for improved speech recognition, and by the IWT project SPACE (sbo/040102): SPeech Algorithms for Clinical an Educational applications.

3. Sensitivity for frame selection

Phonemes differ with respect to manner of articulation, place of articulation, etc., and hence probably also vary in the sensitivity for frame selection. For example, the spectra of some phonemes are relatively stable so they seem suitable for almost random frame elimination. Time evolution of other phonemes is much more critical; therefore these will not be suitable for elimination of any frames, or will be critically dependent on which frames are selected for further processing.

3.1. Normalization

To categorize phonemes based on the sensitivity for frame selection, we adopt a classical statistical hypothesis test by collecting all testing phoneme segments. As illustrated in Fig. 1, a phoneme segment is first normalized to an 11-frame segment either by downsampling or by duplicating some frames, depending on the original number of frames nFr larger or less than 11 frames. The objectives of the normalization are bi-functional: on the one hand it is easy to spot the position of the frame, on the other side a frame in a normalized segment stands for the speech event occurring at its percentage position.

3.2. Sensitivity checking

By either selecting (denoted as 1 in Fig. 1) or not selecting (denoted as 0) a frame from a normalized segment, we can reconstruct this segment in versatile ways. There are $2047(=2^{11}-1)$ possible frame combinations for a normalized segment. For each combination, a simple Viterbi decoder is imposed to recognize the selected frames, resulting in a binary decision, namely 1 if the segment is correctly recognized by the combination, 0 otherwise. For combinations which do not reach the minimum required number of frames (less than three frames), any state is allowed to be the input state or the output state, but the left-to-right decoding topology is kept unchanged. For all segments M of the same phoneme, we collect the results for all combinations and compute their accuracies: $q_i = O_i/M$, where $O_i (1 \leq i \leq 2047)$ is the number of correctly classified segments for the combination i.
In parallel, a classical Viterbi decoder is used as baseline to identify the original segments without frame normalization. The number of correctly recognized segments O follows the binomial distribution $O \sim B(M, c)$, where c is the accuracy achieved by the classical Viterbi decoder. If M is large enough, the normal distribution $N(Mc, Mc(1-c))$ is a good approximation to $B(M, c)$. Therefore, the accuracy c follows $c \sim N(\tilde{c}, \frac{\tilde{c}(1-\tilde{c})}{M})$, where $\tilde{c} = O/M$. A one-side statistical hypothesis test is conducted:

$$H_0 : c = Q; \qquad H_1 : c < Q \tag{1}$$

where $Q = \max_{\forall i}(q_i)$. When for a phoneme the null hypothesis H_0 can not be rejected under a pre-defined significance level α, the phoneme is considered to be insensitive to frame selection. The combination which obtains the highest accuracy Q is denoted as $Bcomb$.

3.3. Good combinations

For frame selection sensitive phonemes, we are more concerned with the combinations $GComb$ which are significantly better than the baseline, but at the same time insignificantly worse than $BComb$. As the shadow part in Fig. 1, the set GS for containing $GComb$ is defined as:

$$GS \rightarrow \{arg(q_i) : q_i \geqslant max(c + \Phi(1-\alpha)\sigma_c, Q - \Phi(1-\alpha)\sigma_Q)\};$$

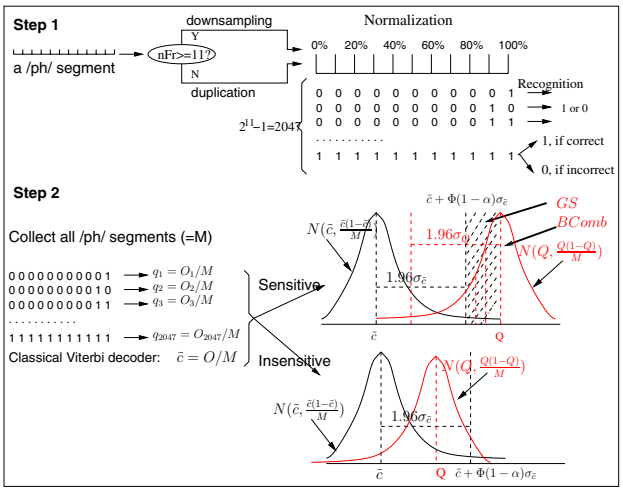


Figure 1: Statistically checking the sensitivity of frame selection

where $\Phi(.)$ is the cumulative distribution function of the normal distribution. The phoneme-dependent combinations $GComb$ in the GS present an approximately identical performance. Analysis of the positions appearing in these combinations reveals the preference of the phoneme. For example, the histogram of the number of selected frames in the GS tells us how many frames are necessary and sufficient to reconstruct the phoneme segments; the probability of the appearance of a certain position in the GS indicates the important parts for the phoneme, etc.
The phoneme-dependent $BComb$, which holds the best performance among all combinations, can be used in a phoneme classification. A difficulty is that according to the $BComb$s of the phonemes, a normalized testing segment is usually sampled to phoneme dependent frame sets. Directly performing recognition over different frame sets is probably misleading. Therefore an alternative decision approach, called $MScore$ is proposed. Suppose the logarithm score of the phoneme dependent frame set $BComb_k$ decoded by the phoneme model $ph$is s_k^{ph}, a testing segment is identified by comparing the sums of the outputs of different phoneme models:

$$ph^* = \underset{\forall ph}{argmax}(\Sigma_{k=1}^{PH} s_k^{ph}). \tag{2}$$

4. Experiment results

4.1. Phoneme classification by frame selection properties

The p-value of the hypothesis test (Eq.1) measures the probability, under the assumption that the H_0 is true, of obtaining a result at least as extreme as Q. The smaller the p-value is, the more likely the frame selection is preferable for the investigated phoneme. Tab. 1 enumerates the list of 44 phonemes as well as their number of segments and their p-values after excluding the silence-like phonemes.
If we set the cut-off significance α as 5%, the 44 phonemes can be categorized as illustrated in Fig. 2. As can be seen, to a great extent the sensitivity to frame selection conforms to natural phonetic classification. For example, most monophthongs tend to prefer the frame selection while most diphthongs do not benefit from the frame selection; for consonants, there is an obvious boundary separating unvoiced phonemes from voiced phonemes. This observation strongly supports our conjecture that the frame selection is only suitable for some phonemes, depending on their intrinsic properties.

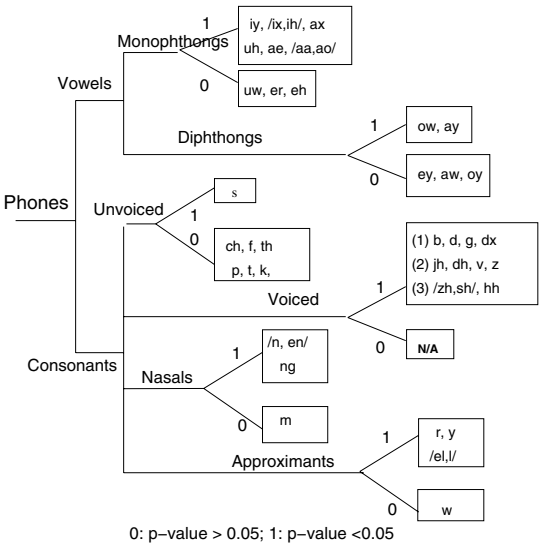


Figure 2: Frame selection sensitivity based classification. Confusions between phoneme pairs within two slashes are not counted [7]

4.1.1. Vowels

Fig. 3 is a graph depicting the relative tongue positions of English vowels [8], plus two glides /y/ and /w/, which are close to the positions of /iy/ and /uw/. The corresponding p-values are also appended. The whole graph is divided into two parts, resulting from the comparison between the p-values and the significance level. We can see that vowels at the high-back corner and at the middle are not influenced much by the frame selection; on the contrary, vowels at the high-front corner or the low part are more sensitive to the frame selection. The three thick arrows show the tendencies towards increasing p-values.
The combinations in the GS are broadly similar both concerning the number of selected frames and concerning the percentage positions they hold. Fig. 4a gives an example for vowel /ae/. Fig. 4 a(1) illustrates the histogram of the number of selected frames in the GS. As can be seen, the numbers of selected frames are concentrated to a certain number, with a gaussian-like shape; selecting too few or too many frames deteriorates the performance. We also plot the probability of the appearance of a certain position in the GS, as shown in Fig. 4 a(2). The higher the probability is, the more likely the speech event occurring at the position is crucial. For instance, for phoneme /ae/, frames at the middle part are frequently selected by the $GComb$s, indicating the importance of this part; but frames at 90% and 100% are almost never preferred.

4.1.2. Consonants

Consonants consist of approximants, nasals, voiced and unvoiced phonemes. Fig. 2 has already shown that the frame selection is favored by almost all vowel-like phonemes, including approximants and nasals except for /w/ and /m/. For the remaining consonants, fricatives and plosives, the separation is based on voicing: all voiced phonemes unanimously accept the frame selection, but all unvoiced phonemes reject it, except for /s/.
Although all voiced phonemes are frame selection sensitive, the required number of frames and frame positions are quite different. Based on these differences, the voiced phonemes can be further divided roughly into three subclasses, as also shown in Fig. 2. The first subclass includes all voiced plosives /b/, /d/, /g/ and /dx/. There are two remarkable characteristics for this

Class	Ph	nSeg	p(%)	Ph	nSeg	p(%)
Monoph-thongs	ax,ah	2295	10^{-8}	uh	215	0.02
	ae	772	10^{-2}	iy	1810	0.3
	ih,ix	3939	0.6	ao,aa	1607	2.9
	er	1692	6.2	eh	1247	9.6
	uw	572	19.5			
Diph-thongs	ow	600	2.9	ay	686	4.5
	aw	216	10.1	ey	802	11.1
	oy	127	22			
Approxi-mants	y	376	0.6	r	1814	0.1
	l,el	2201	1.2	w	903	11.3
Nasals	n,en	2650	10^{-9}	ng	378	1.5
	m	1406	9.0			
Voiceds	dx	634	10^{-4}	g	452	0.5
	d	841	0.9	b	886	1.1
	dh	896	10^{-9}	v	710	10^{-3}
	z	1236	10^{-6}	jh	295	1.4
	zh,sh	533	10^{-5}	hh	561	3.6
Un-voiceds	s	2172	10^{-3}	p	957	21.4
	t	1367	25.0	k	1204	19.4
	ch	259	33.7	f	911	35.7
	th	259	15.9			

Table 1: List of the phonemes and their p-values to the frame selection. nSeg: number of segments in the testing set; p: p-values for the hypothesis tests

subclass. First, the number of selected frames in the combinations of GS is equal to, or even more than the original length. Second, the selected positions are broadly distributed. Fig. 4b illustrates these for phoneme /g/. We can see that most of numbers of selected frames in the GS are concentrated from 6 to 10, considering the average of number of original frames for /g/ is 7.3. Meanwhile the probabilities over all candidate positions are quite high and the distribution is relatively flat. A possible explanation for this phenomenon is that as we conjecture, all parts for a plosive are indispensable, but for voiced plosives, some parts may need to be paid more attention, as we duplicate some frames in the normalization stage. Conversely, the phonemes in the second subclass, consisting of /jh/, /dh/, /v/ and /z/, usually overlook one or more parts, and their required number of frames is significantly less than the original length. The third subclass is composed of the phonemes excluded from the above two subclasses.

4.2. Phoneme classification

In this section, we will make use of the phoneme-dependent $BComb$, to perform a phoneme classification task. A three-fold cross-validation testing scheme is used on the TIMIT testing database: the database is randomly divided into three equal-size subsets, $S0$, $S1$ and $S2$ respectively. In turn, two of them are combined as a development set to achieve the $BComb$s and the remaining subset acts as the real testing set. In the test, a phoneme dependent frame set is obtained from the normalized testing segment according to the $BComb$ of the phoneme. Besides the *MScore* as we introduced in section 3.3, we also present the results of direct decoding, called *AveProb*, which decodes a frame set by the corresponding phoneme model and then the average score per frame is compared over different hypotheses. The essential difference between these two methods is that the *Aveprob* makes the decision based on different frames, but the *MScore* on the same frame set at one time. To investigate whether the sensitivity of frame selection is independent on model structures, the classification rates for more

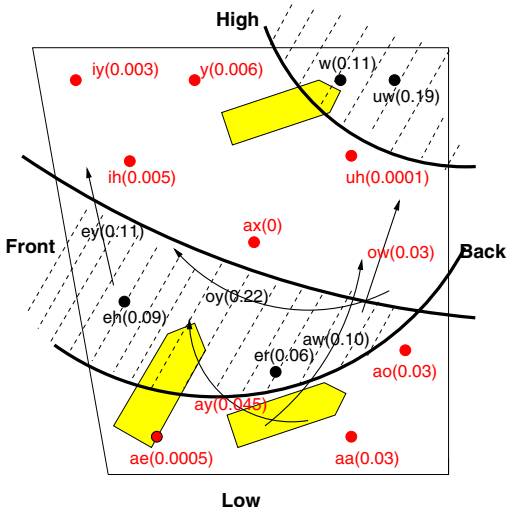


Figure 3: Vowel separation based on frame selection sensitivity

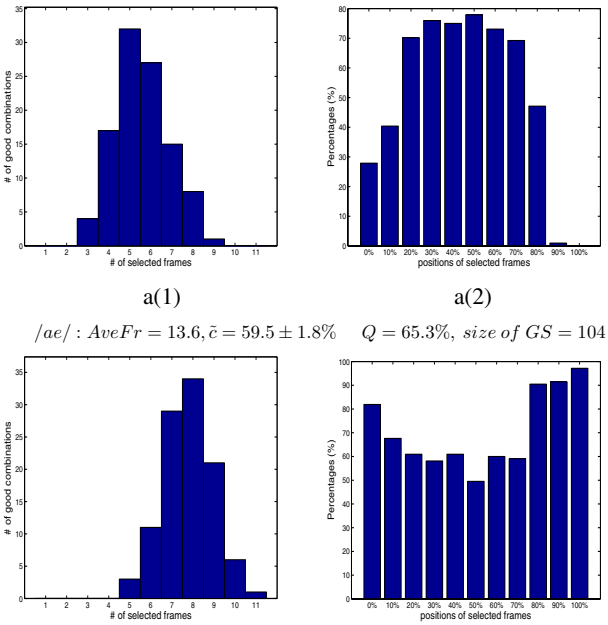
	16 gaussians per state			≈200 gaussians per state		
	Baseline	AveProb	MScore	Baseline	AveProb	MScore
S0	70.0	63.1	**72.0**	77.2	71.0	**77.9**
S1	69.0	63.7	**70.7**	77.1	70.8	**78.1**
S2	69.5	64.0	**71.5**	78.2	72.2	**79.2**
Mean	69.5	63.6	**71.4**	77.5	71.3	**78.4**

Table 2: Phoneme classification rates (%) on TIMIT database. The 95% confidence boundary is $\pm 0.42\%$.

complicated context independent phoneme models are also presented. For these 48 phoneme models, there are 141 HMM states with on average around 200 gaussians per state. Tab. 2 gives the performances on three testing subsets, along with the traditional Viterbi decoder without frame selection using different complexities of HMMs. As can be seen, the *MScore* outperforms the *Baseline* uniformly and significantly, implying that the phoneme-dependent frame position representation indeed reveals an implicit property of phonemes. On the other hand, although *Aveprob* takes the relevant frame positions into account, the performance is largely deteriorated by making decisions on different frame sets.

5. Discussion and conclusions

The phoneme dependent behavior concerning frame selection is quite complex. As we have shown, the decision regions for vowels on the graph of vowel production is somewhat irregular. This irregularity needs further investigation. Consonants show orderly preference to the frame selection. But the behavior of some exceptions, for example /s/, remains unclear so far.
Concerning the frame selection sensitive phonemes, we investigated the number of required frames and the probability a percentage position appears in the GS set. We observed that the combinations in the GS show quite similar distribution, including the length of the combinations and the percentage positions. We also showed that the phoneme-dependent frame selection works successfully in the phoneme classification. Nevertheless it is still difficult to implement the system into a continuous task directly because of the necessity of phoneme boundaries. But it may act as a post-processing stage to re-score the possible hypotheses in the word graph strategy in our future research.



a(1) a(2)

$/ae/ : AveFr = 13.6, \tilde{c} = 59.5 \pm 1.8\%$ $Q = 65.3\%$, $size\ of\ GS = 104$

b(1) b(2)

$/g/ : AveFr = 7.3, \tilde{c} = 63.9 \pm 2.3\%$ $Q = 69.7\%$, $size\ of\ GS = 105$

Figure 4: A vowel example and a consonant example. Left column: the histogram of the number of selected frames for $GComb$; Right column: the probability of a position appearing in the GS

6. References

[1] H. You, Q. Zhu, and A. Alwan, "Entropy-based variable frame rate analysis of speech signals and its application to ASR," in *Proc. ICASSP*, Montreal, Canada, May 2004.

[2] J. Nedel and R. Stern, "Duration normalization for improved recognition of spontaneous and read speech via missing feature methods," in *Proc. ICASSP*, Salt Lake City, USA, May 2001, pp. 313–316.

[3] T.Y. Wu, D. Van Compernolle, J. Duchateau, and H. Van hamme, "Maximum likelihood based temporal frame selection," in *Proc. ICASSP*, Toulouse, France, May 2006, pp. 349–352.

[4] T.Y. Wu, D. Van Compernolle, J. Duchateau, and H. Van hamme, "Single frame selection for phoneme classification," in *Proc. ICSLP*, Pittsburgh, USA, Sept. 2006, pp. 641–644.

[5] Y. Chen and L.-S. Lee, "Energy-based frame selection for reliable feature normalization and transformation in robust speech recognition," in *InterSpeech*, Lisbon, Portugal, Sept. 2005, pp. 385–388.

[6] J. Louradour, K. Daouli, and R Andre-Obrecht, "Discriminative power of transient frames in speaker recognition," in *Proc. ICASSP*, Philadelphia, USA, March 2005, pp. 616–619.

[7] K.F. Lee and H.W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. on ASSP*, vol. 37, no. 11, pp. 1641–1648, 1989.

[8] X. Huang and A. Acero, *Spoken Language Processing*, Prentice Hall, 2001.