A Comparison of Broad Phonetic and Acoustic Units for Noise Robust Segment-Based Phonetic Recognition

Tara N. Sainath and Victor Zue

MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar St. Cambridge, MA 02139, U.S.A
{tsainath, zue}@mit.edu

Abstract

In this paper, we compare speech recognition performance using broad phonetically- and acoustically-motivated units as a pre-processor in designing a novel noise robust landmark detection and segmentation algorithm. We introduce a cluster evaluation method to measure acoustic unit cluster quality. On the noisy TIMIT task, we find that the acoustic and phonetic segmentation approaches offer significant improvements over two baseline methods used in the SUMMIT segment-based speech recognizer, a sinusoidal model method and a spectral change approach. In addition, we find that the acoustic method has much faster computation time in stationary noises, while the phonetic approach is faster in non-stationary noise conditions.

1. Introduction

Nearly all state-of-the-art speech recognition systems employ a sub-word based representation for the mapping between the acoustic signal and words in the lexicon. A sub-word based approach is more effective than a word-based one since the former permits a more parsimonious modeling of context-dependency and enables better sharing of training data. The most commonly used sub-word units are motivated by phonology and phonetics e.g., phonemes, syllables, etc [1]. Phonetic units have the advantage that they are well-defined linguistically, and training of these models is straightforward given the phonetic transcription of an utterance [2]. However, these units may not always be acoustically distinct; consider for example the varying acoustic characteristics throughout a diphthong such as /α^y/. To address this issue, researchers have explored the use of *acoustically*-motivated units [2]. While both of these approaches have been effectively demonstrated for clean speech, we suspect that their performance may vary under conditions where the speech signal has been corrupted by noise. With the increased availability of mobile information devices, we are interested in noise-robust system performance, since speech-based interactions are more likely to be conducted in a wide variety of noisy conditions.

This paper compares speech recognition performance using broad phonetically- and acoustically-motivated modeling approaches. Our investigation is carried out using the SUMMIT speech recognizer, which uses a segment-based framework for acoustic modeling [3]. While a segment-based approach can be formulated as a variable frame-rate Hidden Markov Model (HMM) [4], we suspect the performance of a segment-based system like SUMMIT may be more sensitive to certain types of noise. This is because SUMMIT computes a temporal sequence of frame-based feature vectors from the speech signal, and performs landmark detection based on the spectral energy change of these feature vectors. These landmarks, representing possible transitions between phones, are then connected together to form a graph of possible segmentations of the utterance. While the spectral method works well in clean conditions ([3], [5]), the system has difficulty locating landmarks in noise and often produces poor segmentation hypotheses [6]. In [6], we found that noise robustness in SUMMIT could be improved with a sinusoidal model segmentation approach, which represents speech as a collection of sinusoidal components and detects landmarks from sinusoidal behavior. This method offered improvements over the spectral approach at high signal-to-noise ratios (SNRs), but landmark detection was not as robust at low SNRs.

The purpose of this work is to explore broad phonetically *vs.* acoustically motivated units as a pre-processor to design a noise robust landmark detection method. Specifically, we look at broad classes that are spectrally distinct in noise, such that their transitions occur at large acoustic changes and can aid in landmark detection. Once landmarks are detected, the segment graph is formed and scored similar to the spectral method [3].

First, we introduce a novel cluster evaluation method to choose an appropriate number of acoustic clusters and evaluate their quality. Secondly, on the noisy TIMIT task, we find that our phonetic and acoustic segmentation methods have much lower error rates than the spectral change and sinusoidal methods across all noise types. Finally, we find that the acoustic method has much faster computation time in stationary noises, while the phonetic approach is faster in non-stationary noises.

The remainder of this paper is organized as follows. In Sections 2 and 3, we describe our broad phonetically- and acoustically- derived units, respectively. Section 4 describes our landmark detection and segmentation algorithm using these pre-processors. Section 5 presents the experiments performed, followed by a discussion of the results in Section 6. Finally, Section 7 concludes the paper and discusses future work.

2. Broad Phonetic Units

[1] argues that a phoneme is the smallest phonetic unit in a language to distinguish meaning. Generally phonemes which belong to the same manner class convey similar spectral and temporal properties and can be categorized as belonging to the same broad phonetic class (BPC), while phonemes in different BPCs are acoustically distinct. One representation of these BPCs is vowels, stops, weak fricatives, strong fricatives, nasals and closures [5]. In phonetic classification experiments on the TIMIT corpus [5], it was shown that almost 80% of misclassified phonemes occurred within the same BPC. These BPCs have been shown to be relatively invariant in noise [7], motivating us to define these as our broad phonetic units.

This work was sponsored by the Quanta-MIT T-Party Alliance

3. Broad Acoustic Units

3.1. Learning of Broad Acoustic Units

We learn broad acoustic classes (BACs) from acoustic correlates in the audio signal. The process of learning acoustic units involves a segmentation of the utterance into quasi-stationary sections followed by clustering [2]. We define our segmentation from the underlying phonetic transcription, similar to [8]. Thus, instead of using the underlying phonemes to define BPCs, we learn BACs from acoustic correlates of these phonemes.

First, we cluster the segments to form a set of BACs. To reduce the amount of computation in agglomerative clustering, similar segments are first pre-clustered using an iterative nearest-neighbor procedure to form a set of seed clusters. After the pre-clustering, stepwise-optimal agglomerative clustering is performed on the seed clusters, which merges the two closest clusters at each iteration [8]. This method produces a hierarchical tree-like structure of acoustic clusters, where each level of the tree indicates a different grouping of clusters. After developing a method to learn acoustic structure, it is necessary to evaluate a meaningful number of clusters from the tree-structure.

3.2. Cluster Evaluation with V-Measure

Evaluation measures for supervised clustering methods include *homogeneity* (i.e., purity, entropy), which requires that the clusters contain only data points which are members of a single class, as well as *completeness*, which requires that all data points that are members of a given class are elements of the same cluster. One such recent measure, known as the V-measure [9], derives a clustering metric which evaluates cluster quality by observing the tradeoff between homogeneity and completeness. Evaluation metrics for unsupervised clustering are a bit more difficult, as labels for clusters are not known *a priori*. To evaluate the unsupervised BACs, we slightly alter the V-measure formulation, which we describe below.

Assume we have a set of classes C and clusters K. If the number of clusters K is known *a priori*, the conditional entropy of the classes given the clusters, $H(C|K)$, is defined as:

$$H(C|K) = -\sum_{k=1}^{K} p(k) \sum_{c=1}^{C} p(c|k) \log p(c|k) \qquad (1)$$

Instead of looking at the raw conditional entropy $H(C|K)$, the entropy is normalized by the maximum reduction in entropy the clustering algorithm could provide without any prior cluster information, namely $H(C)$, given by:

$$H(C) = -\sum_{c=1}^{C} p(c) \log p(c) \qquad (2)$$

Using Equations 1 and 2, homogeneity is defined as:

$$homg = \begin{cases} 1 - H(C|K)/H(C) & \text{if } H(C) \neq 0 \\ 1 & \text{if } H(C) = 0 \end{cases}$$

Similarly, completeness is computed by looking at the conditional entropy of the clusters given the classes $H(K|C)$:

$$H(K|C) = -\sum_{c=1}^{C} p(c) \sum_{k=1}^{K} p(k|c) \log p(k|c) \qquad (3)$$

And the worst case value of $H(K|C)$ is $H(K)$, given by:

$$H(K) = -\sum_{k=1}^{K} p(k) log p(k) \qquad (4)$$

Using these metrics, completeness is defined as follows:

$$comp = \begin{cases} 1 - H(K|C)/H(K) & \text{if } H(K) \neq 0 \\ 1 & \text{if } H(K) = 0 \end{cases}$$

The quality of the clustering solution is defined by the V-measure [9], which computes the harmonic mean between homogeneity and complexity as:

$$V_\beta = \frac{(1+\beta) \times homg \times comp}{(\beta \times homg) + comp} \qquad (5)$$

Here β controls the weight for completeness *vs.* homogeneity.

The above V-measure assumes that each class C is labeled. In our work the only labeled classes are the underlying phonemes, and therefore for simplicity we choose these as our classes. However, our goal is to find a set of broad spectrally distinct classes, and using phonemes as classes does not account for this. To address this issue, we assume that cluster k is made up of some true classes c^* which are hidden. Ideally we would like cluster k to be composed of classes which are acoustically similar. We cannot observe these true classes c^*. However, we estimate the distribution $p(c^*|k)$ by the classes our clustering algorithm assigns to cluster k. We can observe the similarity between each of the true classes c^* and all other hypothesized classes c as (i.e., $p(c|c^*, k)$). In addition, we also assume that given c^*, c and k are conditionally independent. Therefore, to calculate $p(c|k)$ we sum over all the hidden variables c^*.

$$p(c|k) = \sum_{c^*} p(c|c^*, k) p(c^*|k) = \sum_{c^*} p(c|c^*) p(c^*|k) \qquad (6)$$

Intuitively, to calculate $p(c|k)$, Equation 6 computes the probability of each of the true classes assigned to cluster k (i.e., $p(c^*|k)$) and weights them by the similarity of these true classes c^* to class c (i.e., $p(c|c^*)$). $p(k|c)$ is computed in the same manner by observing the similarity between c and c^* as:

$$p(k|c) = \sum_{c^*} p(c^*|c, k) p(k|c^*) = \sum_{c^*} p(c^*|c) p(k|c^*) \qquad (7)$$

The confusion probabilities $p(c|c^*)$ and $p(c^*|c)$ are derived from a phonetic classification confusion matrix. Equations 6 and 7 give more weight to classes which are spectrally similar, and Equations 1 and 3 are modified to reflect this as well.

4. Segmentation with Broad Classes

In this section, we discuss how broad classes (i.e., BPCs, BACs), are used to design a robust landmark detection and segmentation algorithm for speech recognition. The spectral change segmentation algorithm [3] hypothesizes landmarks at regions of large spectral change within frame-level feature vectors. More specifically, major landmarks are hypothesized where the spectral change exceeds a specified global threshold. A fixed density of minor landmarks are detected between major landmarks where the spectral change exceeds a specified local threshold. In noisy speech, the system has difficulty locating landmarks due to the static thresholds, resulting in poor segmentation hypotheses. Transitions between spectrally distinct broad classes, generally represent places of largest acoustic change, and thus we explore this as a pre-processer to define landmarks.

Given an input utterance, we first detect the broad classes in the signal. Next, these transitions are used as anchor points for major landmark placement. More specifically, for each broad

class transition, we look at the major landmark setting which best detects the transition while minimizing false alarms. In addition, since each broad class conveys a distinct acoustic characteristic, we look at setting a fixed density of minor landmarks specific to each broad class. For example, stops are more acoustically varying than vowels, and therefore we expect stops to have a greater density of minor landmarks. Finally, major and minor landmarks are connected together to form a segment-based search graph [3]. Full phonetic recognition of the utterance then involves a Viterbi search through the graph.

5. Experiments

Our phonetic recognition experiments are performed on the TIMIT corpus, which offers the benefit of a phonetically-rich context and hand-labeled transcription. We simulate noise on TIMIT by artificially adding pink, speech babble or factory noise, from the Noisex-92 database [10] at SNRs in the range of 30dB to -5dB. We choose these noises because they differ in their stationarity and harmonic properties, allowing us to compare BAC and BPC behavior across different types of noise.

Our experiments explore BPC/BAC units specific to each SNR and noise type. While the number of BPCs is fixed for each condition, the number of BACs vary based on the environment. Each broad class is modeled as a three-state, left to-right context-independent HMM, described in [7]. For a given utterance, broad classes are detected with an HMM, and their transitions are used to aid in landmark detection, as described in Section 4. Phonetic recognition is then done in SUMMIT using triphone acoustic models to score and search the segment graph for the best recognition hypothesis. All broad class and triphone models are trained for each SNR and noise type using the training set. Recognition results are reported on the test set.

First, we analyze the behavior of broad classes in different noises. Secondly, we explore the V-measure for cluster evaluation with and without phonetic similarity. Also, we compare the phonetic error rate (PER) of the BAC and BPC segmentation methods to the baseline sinusoidal and spectral change approaches. Finally, we investigate the recognition computation time, defined as the total time, in seconds, spent during recognition for all test set utterances, for the BAC and BPC approaches.

6. Results

6.1. BPCs vs BACs in Noise

To understand the behavior of broad classes in noise, we analyze the confusion of vowels and fricatives with other phonemes. Figure 1(a) shows the confusions for vowels in each noise type and SNR, normalized by the maximum vowel confusions over all noises. Notice that vowels have the least amount of confusions in stationary, non-harmonic pink noise, implying that harmonics are well-preserved in pink compared to non-stationary, non-harmonic babble, which has the most number of confusions. Non-stationary, non-harmonic factory noise retains harmonics better than babble but not as well as pink. Figure 1(b) plots the normalized confusions for fricatives, and indicates that fricatives have a common amount of confusions and thus behave similarly in all noises. This same trend is true for other non-harmonic broad classes such as stops and closures.

6.2. V-Measure For Choosing Clusters

In this section, we discuss how the V-measure allows us to choose an optimal number of clusters, and to identify the benefits of our novel phonetic similarity measure for better cluster selection at lower SNRs. Figure 2 shows the V-measure with and without the phonetic similarity measure for (a) 30dB and (b) 10dB of babble noise as the number of clusters is varied.

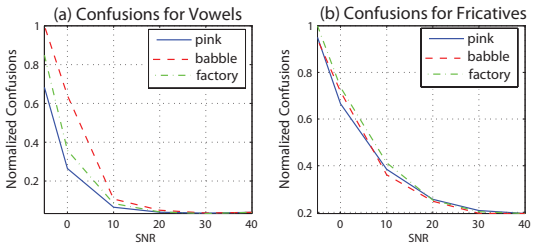

Figure 1: Normalized Confusions in Noise

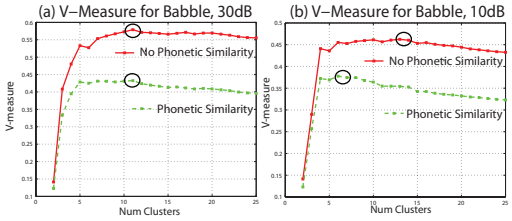

Figure 2: V-measure Metrics vs. Number of Clusters

First, both measures show a broad peak, which represents the best tradeoff between completeness and homogeneity. This peak defines the range for the optimal cluster number. In plot (a), both metrics peak at 11 clusters. In plot (b), the phonetic similarity V-measure peaks at 6, while the other condition peaks at 13. As the SNR decreases and the number of confusions between broad classes increases, as illustrated in Figure 1, intuitively the number of broad classes should decrease. While the V-measure without phonetic similarity seems to find a reasonable number of clusters at 30dB, the increase in clusters at 10dB indicates the clusters are not acoustically distinct. However, using similarity information results in clusters chosen based on spectral closeness, as reflected by a decrease in the number of clusters with decreasing SNR. While only babble is shown here, similar V-measure trends were observed for other noise types.

6.3. Segmentation Error Rates

Table 1 shows the PER for each SNR, averaged across the three noises, for the BPC, BAC, sinusoidal and spectral change methods, while Figure 3 shows the average duration difference between true phoneme boundaries and hypothesized landmarks for each methods. First, decreasing the SNR results in rapid degradation in performance for the spectral change method, as well as a large time deviation from the true phonetic boundaries. While the sinusoidal model approach is more robust at lower SNRs than the spectral change method, it does not perform as well at high SNRs, as landmarks are not as robust. The BAC and BPC methods provide the best performance of all methods, and have the most robust landmarks, as shown in Figure 3. The only exception to this is -5dB of babble noise, where harmonics are very poorly preserved, leading to poor BACs. While the performance of these two methods are fairly similar across noises, Section 6.4 will show that their computation times are different.

6.4. Segmentation Computation Time

In this section, we use the V-measure to investigate the quality of the hypothesized BPC and BAC units, and show the direct correlation to computation time. To assign a set of labeled

TIMIT Average Phonetic Error Rates				
db	bpc	bac	sine	spec
Clean	27.7	**27.3**	30.6	28.7
30dB	28.4	**28.3**	31.3	29.2
20dB	**31.5**	31.7	34.3	32.5
10dB	41.1	**40.9**	43.4	42.1
0dB	**57.9**	58.0	59.4	70.7
-5dB	**67.3**	69.6	68.5	91.8
Average	**42.3**	42.6	45.5	49.2

Table 1: PERs for Segmentation Methods on TIMIT

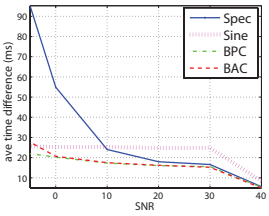

Figure 3: Ave. Time Diff. for True Phonemes and Hyp. Lmks.

classes to the broad units to compute the V-measure, we look at the true underlying phonemes which make up the different BPCs or BACs generated from recognition hypotheses. Figure 4 shows the total V-measure, average V-measure for vowels, and computation time (CPU Time) as a percentage of real time, for the BAC/BPC units in the three noise conditions.

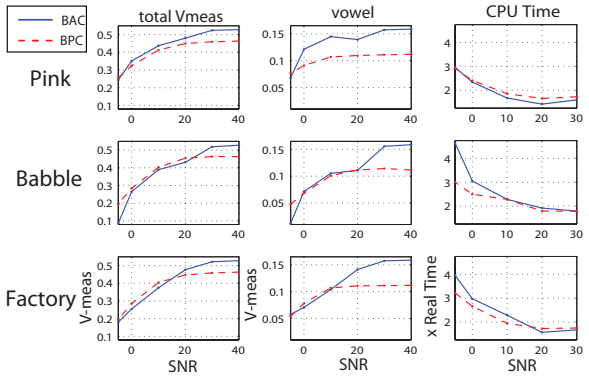

Figure 4: V-measures and CPU Times for Different Noises

In pink noise, the total V-measure is higher for the BAC method across all SNRs, and gains are made particularly in the vowel class. As discussed in Section 6.1, pink noise tends to preserve harmonics well, resulting in a higher V-measure and better quality clusters for the BAC relative to BPC, which groups all vowels into one class. This leads to a faster CPU time for the BAC method. The segment graph in Figure 5(a) also shows that the BAC method has more finer level hypothesized acoustic clusters compared to the BPC method in Figure 5(b), resulting in a smaller segment graph and faster CPU time.

In babble noise, harmonics are not well preserved at lower SNRs. This leads to greater confusions between broad classes, resulting in fewer BACs. Thus, in babble the BPC method has a higher V-measure and faster CPU time at lower SNRs.

Finally, for factory noise, at high SNRs, harmonics are well-preserved and the BAC method has a higher V-measure and faster CPU time. As the SNR decreases, harmonics are not as well preserved in factory compared to pink and the number of BACs decreases. Thus, the BPC method has finer level BPCs and is faster at lower SNRs, as confirmed by the smaller segment graph for BPC in Figure 5(c) compared to BAC in 5(d).

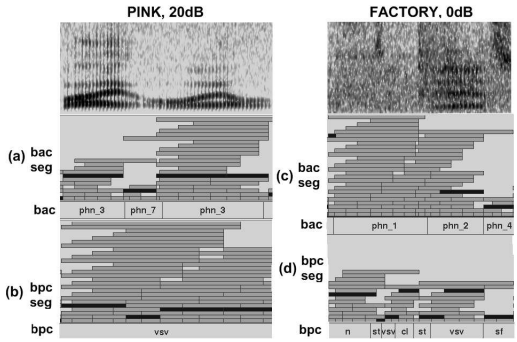

Figure 5: Graphical displays of BAC and BPC methods in SUMMIT. The top display contains speech spectrograms. Below that, (a) shows a segment-network for the BAC method in pink noise, and *bac* indicates the hypothesized BACs. Similarly, (b) shows the network for the BPC method in pink, and *bpc* are the hypothesized BPCs. The darker colored segments indicate the highest scoring segmentation achieved during search. (c) and (d) show the BAC and BPC methods in factory noise.

7. Conclusions

In this paper, we explored BPCs and BACs under different noise conditions in designing a robust segment-based algorithm. We showed our novel segmentation algorithm outperformed the baseline spectral change and sinusoidal methods. Also, we introduced a phonetic similarity metric into the V-measure, which allowed us to choose an appropriate number of distinct acoustic clusters and analyze under what noises the BAC or BPC method is preferred. We found that the BPC method has faster CPU time in non-stationary noises, while BAC is faster in stationary conditions. Finally, preliminary results on the Aurora-2 noisy digit task using the BAC and BPC segmentation approaches indicate the best results on a segment-based recognizer to date and suggest the generalizability of these methods to other tasks.

8. References

[1] G. Fant, *Acoustic Theory of Speech Production*, Mouton and Co., 's-Gravenage, Netherlands, 1960.

[2] C. Lee, F.K. Soong, and B. Juang, "A Segment Model Based Approach to Speech Recognition," in *ICASSP*, New York, 1988.

[3] J. R. Glass, "A Probabilistic Framework for Segment-Based Speech Recognition," *Computer Speech and Language*, 2003.

[4] M. Ostendorf, V. Digalakis, and O. Kimball, "From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition," *IEEE Transactions on Speech and Audio Proceesing*, vol. 4, no. 5, pp. 360–378, September 1996.

[5] A. Halberstadt and J. Glass, "Heterogeneous Acoustic Measurements for Phonetic Classification," in *Eurospeech*, 1997.

[6] T. N. Sainath and T. J. Hazen, "A Sinusoidal Model Approach to Acoustic Landmark Detection and Segmentation for Robust Segment-Based Speech Recognition," in *Proc. ICASSP*, 2006.

[7] T. N. Sainath, D. Kanevsky, and B. Ramabhadran, "Broad Phonetic Class Recognition in a Hidden Markov Model Framework using EBW Transformations," in *Proc. ASRU*, 2007, pp. 306–311.

[8] J. Glass, *Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition*, Ph.D. thesis, MIT, 1988.

[9] A. Rosenberg and J. Hirschberg, "V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure," in *Proc. EMNLP*, 2007.

[10] A. Varga et. al, "The Noisex-92 Study on the Effect of Additive Noise on Automatic Speech Recognition," Tech. Rep., DRA Speech Research Unit, 1992.