SIGNIFICANCE OF INVARIANT ACOUSTIC CUES IN A PROBABILISTIC FRAMEWORK FOR LANDMARK-BASED SPEECH RECOGNITION

Amit Juneja & Carol Espy-Wilson

Department of Electrical and Computer Engineering, Institute for Systems Research
University of Maryland, College Park, MD 20742
juneja@glue.umd.edu

ABSTRACT

A probabilistic framework for landmark-based speech recognition that utilizes the sufficiency and context invariance properties of acoustic cues for phonetic features is presented. Binary classifiers of the manner phonetic features "sonorant", "continuant" and "syllabic" operate on each frame of speech, each using a small number of relevant and sufficient acoustic parameters to generate probabilistic landmark sequences. The relative nature of the parameters developed for the extraction of acoustic cues for manner phonetic features makes them "invariant" of the manner of neighboring speech frames. This invariance of manner acoustic cues makes the use of only those three classifiers along with the speech/silence classifier complete irrespective of the manner context. The obtained landmarks are then used to extract relevant acoustic cues to make probabilistic binary decisions for the place and voicing phonetic features. Similar to the invariance property of the manner acoustic cues, the acoustic cues for place phonetic features extracted using manner landmarks are invariant of the place of neighboring sounds. Pronunciation models based on phonetic features are used to constrain the landmark sequences and to narrow the classification of place and voicing. Preliminary results have been obtained for manner recognition and the corresponding landmarks. Using classifiers trained from the phonetically rich TIMIT database, 80.2% accuracy was obtained for broad class recognition of the isolated digits in the TIDIGITS database which compares well with the accuracies of 74.8% and 81.0% obtained by a hidden Markov model (HMM) based system using mel-frequency cepstral coefficients (MFCCs) and knowledge-based parameters, respectively.

INTRODUCTION

A probabilistic framework for a landmark-based approach to speech recognition based on representation of speech sounds by bundles of binary-valued phonetic features (Chomsky and Halle, 1968) is presented. The framework exploits two properties of the developed acoustic cues of distinctive features – sufficiency and invariance. Sufficiency of a small number of acoustic parameters (APs) that target the acoustic correlates of a phonetic feature makes the framework use only those APs for a probabilistic decision on that feature. Invariance of APs for a phonetic feature is assumed (and verified in this work) with the variation of context, for example, the APs for the feature sonorant are assumed to be invariant of whether the sonorant frame is in a vowel, nasal or a semivowel context. Similarly, the APs for the place feature alveolar of stop consonants are assumed to be independent of the vowel context (Stevens, 1999). In this paper, it is shown how the framework utilizes the two properties of the APs.

Although the APs may not strictly possess these properties, it is shown using the APs for one of the phonetic features that these properties may be approximately correct.

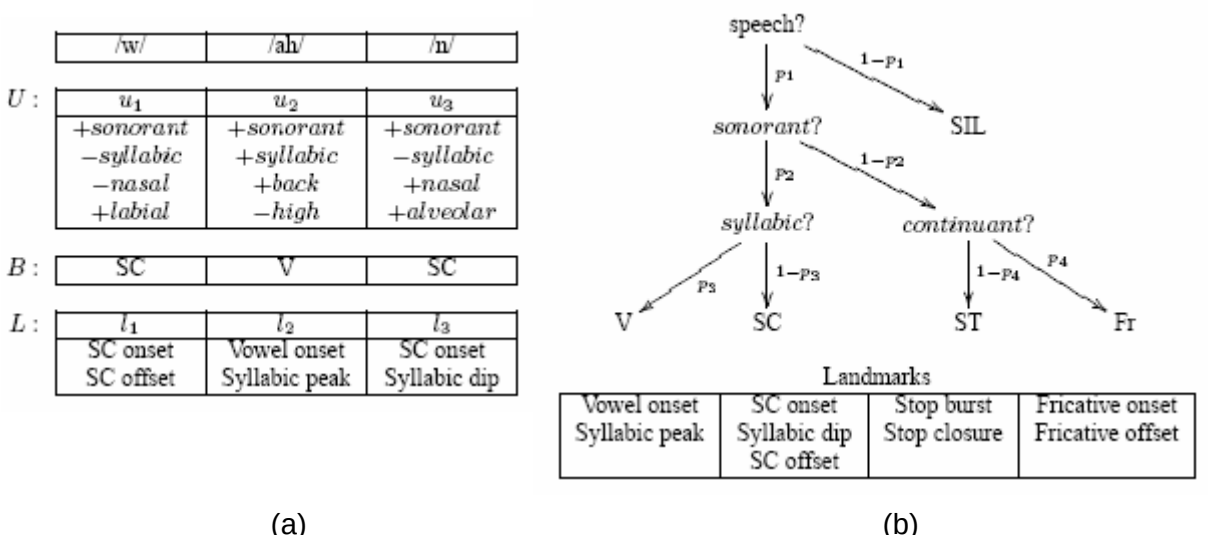
In the enhancements to the event-based system (EBS) (Espy-Wilson, 1994, Bitar, 1997) presented in this article, probabilistic landmark sequences related to manner phonetic features are located in the speech signal. The landmarks are then analyzed for place and voicing phonetic features, resulting in a complete set of features for the description of a word or a sentence. The lexicon can then be accessed using the representation of words in terms of sequences of bundles of phonetic features. By selectively using knowledge based APs and conducting landmark-based analysis, EBS offers a number of advantages. First, the analysis at different landmarks can be carried out by a different procedure, e.g., a higher resolution can be used in the analysis of burst landmarks of stop consonants than for the syllabic peak landmarks of vowels. Second, a different set of measurements can be used at different landmarks for the analysis of the place features. Third, the selective analysis makes it easy to pinpoint the exact source of errors in the recognition system.

The presented probabilistic framework for EBS is similar to the SUMMIT framework (Halberstadt, 1998) in the sense that both systems carry out multiple segmentations and then use these segmentations for further analysis. Unlike the SUMMIT system, EBS is strictly based on acoustic landmarks and articulatory features and uses the idea of minimal knowledge based acoustic correlates of phonetic features. EBS requires only binary classifiers operating on a fixed number of parameters for each classifier, obtained from each frame of speech in the case of manner classification and from specific landmarks in the case of place and voicing classification. Support Vector Machines (SVMs) (Vapnik, 1995) have been chosen for the purpose of classification because SVMs have been shown to be effective for distinctive feature detection in speech (Niyogi, 1998).

METHOD

The probabilistic framework presented in this section assumes the sufficiency and invariance properties of the APs. The validity of these assumptions is then discussed in the next section. The problem of speech recognition can be expressed as maximization of the posterior probability $P(LU \mid O) = P(L \mid O)P(U \mid LO)$ of landmark sequence $L = \{l_i\}_{i=1}^{M}$ and the sequence of feature bundles $U = \{U_i\}_{i=1}^{N}$, given the observation sequence O. The meaning of these symbols is explained in Figure 1(a) which shows their canonical values for the word “one”. Landmarks can be associated with five broad classes – vowel (V), fricative (Fr), sonorant consonant (SC), stop burst (ST) and silence (SIL) – as shown in Figure 1(b). The sequence of landmarks for an utterance can be completely determined by its broad class sequence. Therefore, $\mathrm{P(L \mid O) = P(B \mid O)}$ where $\mathrm{B} = \{\mathrm{B_i}\}_{i=1}^{M}$ is a sequence of broad classes for which the landmark sequence L is obtained. Note that there is no temporal information contained in B, L and U and these are only sequences of symbols.

Given a sequence O ={o_1, o_2,…, o_T,} of T frames, where o_t is the vector of all APs at time t, the broad class segmentation problem can be stated as a maximization of $P(BD \mid O)$ over all B and D. EBS does not use all of the APs at each frame; however, all of the APs are assumed to be available so as to develop the probabilistic framework. EBS uses the probabilistic phonetic



(a) (b)

Figure 1. (a) Illustration of symbols B, L and U, (b) Probabilistic phonetic feature hierarchy and landmarks

feature hierarchy in Figure 1(b) to segment speech into the four broad classes and silence. The concept of probabilistic hierarchies has appeared before with application to phonetic classification (Halberstadt, 1998), but, to the best of our knowledge, it has not been used as a uniform framework for landmark detection and phonetic classification. Calculation of $P(BD \mid O)$ for all D is computationally very intensive. Therefore, $P(B \mid O)$ is approximated as $\max_D P(BD \mid O)$. This approximation is similar to the one made by the Viterbi algorithm in HMM decoding as well as by the SUMMIT system (Halberstadt, 1998). Using the probabilistic hierarchy, the posterior probability of a frame being part of a vowel at time t can be expressed as (using P_t to denote the posterior probability of a feature or a set of features at time t)

(1)

$$P_t(V \mid O) = P_t(speech, sonorant, syllabic \mid O)$$
$$= P_t(speech \mid O) P_t(sonorant \mid speech, O) P_t(syllabic \mid speech, sonorant, O)$$

Similar expressions can be obtained for the rest of the manner classes. To calculate the posterior probability of a manner phonetic feature at time t, we only need to pick the acoustic correlates of the feature in a set of frames $\{t-s, t-s+1, \ldots, t+e\}$ using s previous frames and e following frames. Let this set of acoustic correlates extracted from the analysis frame and the adjoining frames for a feature f be denoted by x_t^f. Then, applying the sufficiency property of acoustic correlates,

(2)

$$P_t(V \mid O)$$
$$= P_t(speech \mid x_t^{speech}) P_t(sonorant \mid speech, x_t^{sonorant}) P_t(syllabic \mid speech, sonorant, x_t^{syllabic})$$

The probability $P(BD|O)$ can now be expanded in terms of the underlying manner features of each broad class. Denote the features for class B_i as $\{f_1^i, f_2^i, ..., f_{N_{B_i}}^i\}$, the broad class at time t as b_t and the sequence $\{b_1, b_2, ..., b_{t-1}\}$ as b^{t-1}. Now making a stronger use of sufficiency of APs even when b^{t-1} is given,

(3)

$$P(B,D|O) = \prod_{i=1}^{M} \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} P(B_i|O,b^{t-1}) = \prod_{i=1}^{M} \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} \prod_{k=1}^{N_{B_i}} P_t(f_k^i | x_t^{f_k^i}, f_1^i, f_2^i, ..., f_{k-1}^i, b^{t-1})$$

where the right side is the multiplication of posterior probabilities of manner phonetic features given (a) the APs of those features, (b) the preceding broad classes b^{t-1}, and (c) manner features above the current feature in the hierarchy. Now x_t^f is assumed as independent from b^{t-1} given the two sets - $\{f_1^i, f_2^i, ..., f_k^i\}$ and $\{f_1^i, f_2^i, ..., f_{k-1}^i\}$. The independence or invariance given the first set can be shown to be approximately true for APs (next section). Applying these assumptions,

(4)

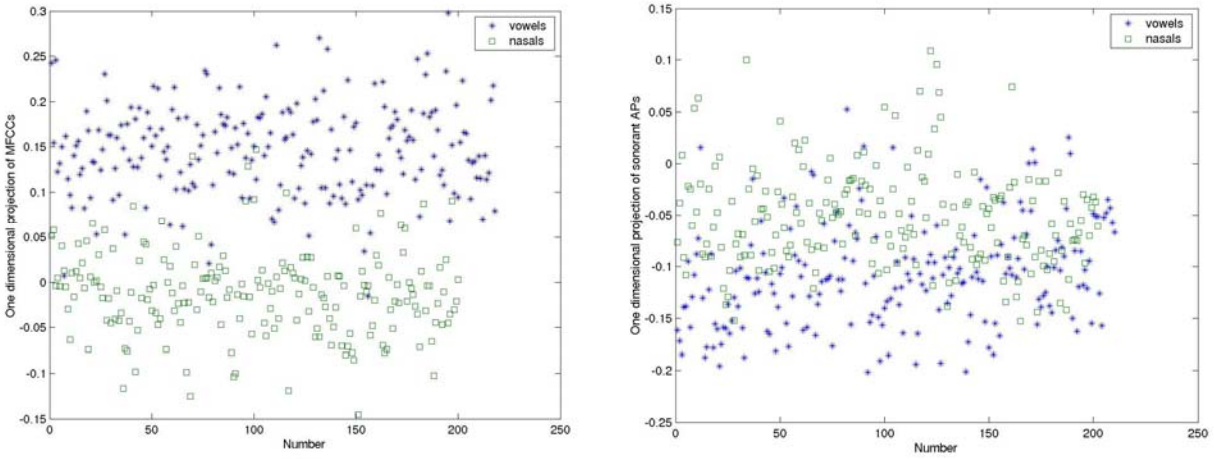
$$P(B,D|O) = P(B) \prod_{i=1}^{M} P(D_i|B_i) \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} \prod_{k=1}^{N_{B_i}} \frac{P_t(f_k^i | x_t^{f_k^i}, f_1^i, f_2^i, ..., f_{k-1}^i)}{P_t(f_k^i | f_1^i, f_2^i, ..., f_{k-1}^i)}$$

where the duration of a broad class segment is assumed to be dependent only on that broad class. The posterior probabilities in the numerator on the right side are computed using the outputs of SVMs, where one SVM is trained for each feature, and the conversion of SVM outputs to probabilities is done using binning (Drish 1998). Duration is modeled using a mixture of Rayleigh densities and the parameters of the distribution are found empirically from the training data. An N-best probabilistic segmentation algorithm similar to (Lee, 1998) is used to compute $P(B|O)$ for different B. The segmentation provides all the landmarks except the syllabic peaks and dips which can be found by using the minima or maxima of energy in (640Hz-2800Hz). Segmentation paths and hence the landmark sequences can be constrained using a broad class pronunciation graph. An expression for $P(U|OL)$ can be derived using the sufficiency and invariance properties of APs for place, voicing and nasal features (Juneja, 2003).

SUFFICIENCY AND INVARIANCE

Although it is not clear how these properties can be rigorously established for certain parameters, some idea can be obtained from classification and scatter plot experiments. For example, sufficiency of the four APs used for sonorant feature detection – periodicity, aperiodicity, energy in (100Hz,400Hz) and ratio of the energy in (0,F3) to the energy in (F3,sampling rate) – can be viewed in relation to 13 MFCCs in terms of classification accuracy of the sonorant feature. Using SVMs, a frame classification accuracy of 94.39% was obtained on TIMIT ‘sx’ sentences which compares well to 94.68% accuracy obtained using MFCCs, when

all other test conditions were kept identical. Similar results for other phonetic features have been obtained (Juneja and Espy-Wilson, in preparation).



(a) (b)

Figure 2. (a) Projection of 13 MFCCs into a one-dimensional space with vowels and nasals as discriminating classes, (b) Similar projection for four APs used to distinguish +sonorant sounds from –sonorant sounds.

Invariance was assumed in Equation 4 where APs for a manner feature x_t^f were assumed to be independent of the manner class labels of preceding frames. A typical case where this may not be true is when the APs for the sonorant feature are assumed to be invariant of whether the analysis frame lies in the middle of a vowel region or the middle of a nasal region. Such independence can roughly be measured by the similarity in the distribution of vowels and nasals based on the APs for the feature sonorant. Figure 2(a) shows projection of the 13 MFCCs for the sonorant feature into a one-dimensional space using Fischer Linear Discriminant Analysis (LDA). A similar projection is shown for the four APs in Figure 2(b). It can be seen from these figures that there is considerably more overlap in the scatter of the vowels and the nasals for the APs of the sonorant feature than for the MFCCs. Thus, the APs for the sonorant feature are more independent of the manner context than are the MFCCs.

RESULTS

Results are shown in Table 1 for broad class recognition on all the 804 test ‘sx’ sentences of the TIMIT database as well as isolated digits from TIDIGITS database (Juneja and Espy-Wilson, in preparation). The ‘si’ sentences from the TIMIT training database were used for training. Results are compared to two HMM-based systems – one that used APs and the other that used 39 cepstrum-based parameters. The performance of all the three systems is comparable when testing on TIMIT. The performance of HMM-MFCC system drops significantly when using the models trained on TIMIT for unconstrained broad class recognition on TIDIGITS, in comparison to HMM-AP and EBS. This is in agreement with the speaker independence of APs shown earlier (Deshmukh et al, 2002). But the result that is most relevant in the current context is that the

performance of EBS compares well to the HMM-AP system in all the three testing conditions, though EBS uses a maximum of 5 APs for any decision while HMM-AP system uses all the APs for all the models.

Table 1. Results (Correct/Accurate in percent, Tr: training, Te: Testing)

	EBS	HMM-AP	HMM-MFCC
Tr: TIMIT, Te: TIMIT	86.7/79.5	83.5/78.2	86.6/80.4
Tr: TIMIT, Te: TIDIGITS (Unconstrained)	91.5/80.2	87.9/81.0	88.3/74.8
Tr: TIMIT, Te: TIDIGITS (Constrained)	90.8/85.0	90.3/85.3	92.3/84.2

CONCLUSION

A probabilistic system for detection of acoustic landmarks using binary classifiers of manner features performs comparable to an HMM based system. The framework uses two properties of APs – sufficiency and context invariance. The knowledge based APs have been shown to satisfy these properties more than the cepstrum based coefficients. The probabilistic framework formalizes the requirement for development of invariant acoustic cues for speech recognition.

ACKNOWLEDGEMENTS

This work was supported by NSF grant #BCS-0236707 and Honda Initiation Grant 2003.

REFERENCES

Bitar, N., (1997) Acoustic analysis and modeling of speech based on phonetic features, Ph.D. thesis, Boston University.

Chomsky, N., and N. Halle, (1968) The Sound Pattern of English, Harper and Row.

Deshmukh, O., C. Espy-Wilson, and A. Juneja, (2002) Acoustic-phonetic speech parameters for speaker independent speech recognition, ICASSP.

Drish, J., (1998) Obtaining calibrated probability estimates from support vector machines, http://citeseer.nj.nec.com/drish01obtaining.html.

Espy-Wilson, C., (1994) A feature-based semivowel recognition system, JASA, 96 , 65–72,

Halberstadt, A. K., (1998) Heterogeneous acoustic measurements and multiple classifiers for speech recognition, Ph.D. thesis, MIT.

Juneja, A. and Espy-Wilson, C. (in preparation), Probabilistic detection of acoustic landmarks using acoustic-phonetic information, Journal of the Acoustical Society of America.

Juneja, A., (2003) Speech recognition using acoustic landmarks and binary phonetic feature classifiers, PhD Thesis Proposal, University of Maryland College Park, http://www.ece.umd.edu/~juneja/proposal.pdf

Lee, S., (1998) Probabilistic segmentation for segment-based speech recognition, Master's thesis, MIT.

Niyogi, P., (1998), Distinctive feature detection using support vector machines, ICASSP

Stevens, K.N., and Manuel, S. Y., (1999) Revisiting Place of Articulation Measures for Stop Consonants: Implications for Models of Consonant Production, ICPhS