Landmark-based Approach to Speech Recognition: An Alternative to HMMs

Carol Y. Espy-Wilson[1], *Tarun Pruthi*[1], *Amit Juneja*[2], *Om Deshmukh*[1]

[1]Institute for Systems Research and Dept. of Electrical and Computer Eng.
University of Maryland, College Park, MD 20742, USA
[2]Think-A-Move, Ltd., Beachwood, OH 44122, USA
espy@umd.edu, tpruthi@umd.edu, amjuneja@gmail.com, omdesh@umd.edu

10.21437/Interspeech.2007-324

Abstract

In this paper, we compare a Probabilistic Landmark-Based speech recognition System (LBS) which uses Knowledge-based Acoustic Parameters (APs) as the front-end with an HMM-based recognition system that uses the Mel-Frequency Cepstral Coefficients as its front end. The advantages of LBS based on APs are (1) the APs are normalized for extra-linguistic information, (2) acoustic analysis at different landmarks may be performed with different resolutions and with different APs, (3) LBS outputs multiple acoustic landmark sequences that signal perceptually significant regions in the speech signal, (4) it may be easier to port this system to another language since the phonetic features captured by the APs are universal, and (5) LBS can be used as a tool for uncovering and subsequently understanding variability. LBS also has a probabilistic framework that can be combined with pronunciation and language models in order to make it more scalable to large vocabulary recognition tasks.

Index Terms: landmark, speech recognition, acoustic parameters, phonetic features.

1. Introduction

State-of-the-art ASR systems are based on Hidden Markov Modeling (HMM) and the standard parameterization of the speech signal consists of Mel-Frequency Cepstral Coefficients (MFCCs) and their first and second derivatives [1, 2]. The HMM framework assumes independence of the speech frames so that each one is analyzed and all of the MFCCs are looked at in every frame. In contrast, a landmark-based approach to speech recognition targets level of effort where it is needed. The extraction of relevant information is performed in two steps. First, landmarks that signal significant articulatory changes (e.g., changes in the manner of articulation, a sudden release of air pressure and changes in the state of the larynx) are detected. This step may involve the analysis of each speech frame. Second, further analysis is carried out only in specific regions designated by the landmarks to extract other relevant acoustic information regarding place of articulation to help in the classification of the sounds spoken. This two-step process in effect takes into account the strong correlation among the speech frames so that every frame is not always analyzed.

In addition to being more efficient, the landmark approach is more flexible. Analysis at different landmarks can be done with different resolutions. For example, the transient burst of a stop consonant may be only 5 ms long. Thus, a short temporal window is needed for analysis. On the other hand, vowels which are considerably longer (50 ms for a /schwa/ to 300 ms for an /ae/) need a longer analysis window. Another important feature is that the Acoustic Parameters (APs) used to extract relevant information will depend upon the type of landmark. For example, at a burst landmark, appropriate APs will be those that characterize the spectral shape of the burst (maybe relative to the vowel to take into account contextual influences) to distinguish between labial, alveolar and velar stops. However, at a vowel landmark, appropriate APs will be those that look at the relative spacing of the first three formants to determine where the vowel fits in terms of the *phonetic features* (minimal binary valued units that are sufficient to describe all the speech sounds in any language [3]) front, back, high and low. Given the physical significance of the APs and a recognition framework that uses only the relevant APs, error analysis often points to variability that has not been accounted for. Further, given the extensive variability in the speech signal, a complete landmark-based system would integrate this front end processing with a lexical access system that handles pronunciation variability and takes into account prosody, grammar, syntax and other higher-level information. The probabilistic framework which has been developed for this system allows the integration of higher level information in a systematic manner.

The rest of the paper is organized as follows: Section 2 gives a detailed description of the Landmark-Based System (LBS) being developed in our lab. This includes a detailed description of the front-end of APs which are based on the knowledge of the human speech production system, and the probabilistic framework which acts as the back-end. This section also presents results that have been obtained until now for the different components of this system. Section 3 summarizes the main points of the paper.

2. Landmark-based System

2.1. Knowledge-based Acoustic Parameters

The front-end processing in this system includes the extraction of APs that are based on the knowledge of human speech production system. APs are exact measures extracted from the speech signal or its time-frequency representation that signal acoustic correlates of phonetic features. These features encode universal decision boundaries in the nonlinear mapping between speech production and speech perception [4]. The 20 or so features characterizing all of the world's languages can be divided into three sets: (1) manner-of-articulation features that relate to how open or closed the vocal tract is, (2) place-of-articulation features that specify where the main constriction is located in the vocal tract, and (3) source features that specify the opening of the glottis and vibration of the vocal folds. Speech perception results beginning with those in [5] have demonstrated two important facts about phonetic features. First, across a wide

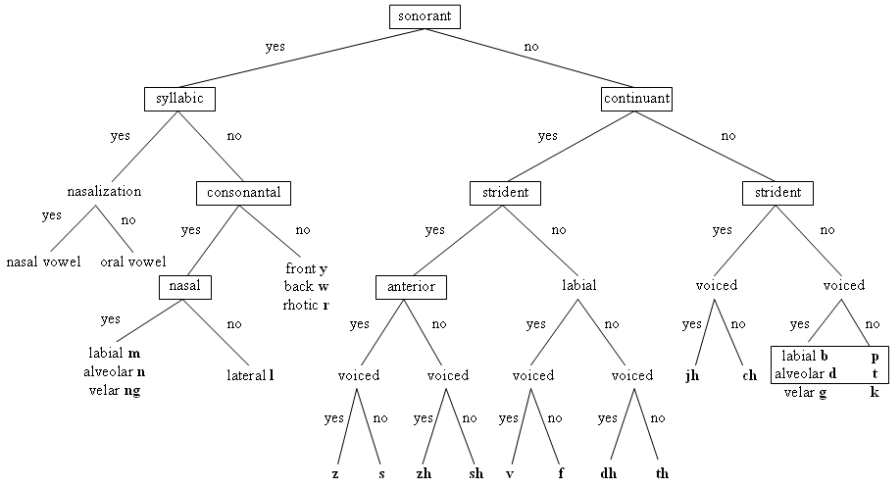

Figure 1: Phonetic Feature Hierarchy. APs exist for the boxed features.

Table 1: Digit recognition accuracy (%) (TI46 corpus for Adults, TIDIGITS corpus for children). Ad = Adult, Ch = Child, Fe = Female, Ma = Male.

Train/Test	Ad/Ad	Fe/Ma	Ma/Fe	Ch/Ch	Ad/Ch	Ch/Ad
39 MFCCs	99.88	68.29	70.27	98.30	60.20	62.37
30 APs	99.53	79.24	90.90	97.50	85.70	89.81

range of SNRs and channel conditions, the probability that a listener will mistakenly hear one phoneme in place of another may be partitioned into independent factors corresponding to individual phonetic feature errors, i.e., errors in the perception of individual features are independent. Second, phonetic features can be hierarchically arranged based on differential sensitivity to SNR. Figure 1 shows the features arranged in a hierarchical tree-structure, with the manner features at the upper levels, the place and voicing features at the lower levels and the different speech sounds at the terminal nodes. This structure reflects the features differential sensitivity as first proposed in [5], with those at higher levels showing less sensitivity to SNR than those at lower levels. The hierarchy shows how the phonetic features can be combined to group sounds into classes such as vowel (+sonorant, +syllabic), fricative (-sonorant, +continuant) and stop (-sonorant, -continuant). Further, it reflects how phonetic features can be used to distinguish between speech sounds, e.g, /z/ and /s/ are distinguished by the feature voiced.

2.1.1. Speaker Invariance

Our philosophy in developing APs that reflect the phonetic structure of language is that they be based on measures that normalize for extra-linguistic information, such as individual differences among speakers and channel characteristics. At present, we have developed 30 APs for 14 of the 20 or so phonetic features (Figure 1. We have conducted several experiments to compare our signal representation so far with the 39 MFCCs commonly used in ASR front end systems (13 MFCCs+d+dd, cepstral mean subtraction, implemented in the hidden Markov model toolkit HTK) [2]. Table 1 summarizes our results across databases and speaker groups. These results had earlier been presented in [6, 7]. These results clearly show that the knowledge-based APs are much more robust to cross-gender and cross-age tests.

2.1.2. Language Invariance

It has been suggested in past literature that the proposed phonetic features are universal in the sense that they can represent all the sounds in all languages of the world [3]. Thus, with this approach the effort to move from one language to another can be reduced if reliable algorithms can be developed to extract APs for these phonetic features which work across languages.

This idea of the universality of features was tested in the context of the APs for vowel nasalization (see [8] for details of the APs). In this experiment, a Linear SVM classifier was trained to distinguish between oral and nasalized vowels using a database of American English, WS96/97, a part of the switchboard telephone speech corpus [9], and tested on the same task on the test set of WS96/97, and a database of Hindi Language called OGI Multilanguage telephone speech corpus [10]. Table 2 shows a comparison of the results obtained on this cross-language task using the proposed knowledge-based APs. For comparison purposes, results obtained by using MFCCs in the same experimental framework are also shown in this table. These results not only show that the performance of the knowledge-based APs is much better than the performance of the MFCCs on this task, but also show that the APs generalize much better when tested on a cross-language task. It must be noted that the performance obtained with MFCCs for the correct classification of nasalized vowels is even below the chance accuracy of 50%. Thus, it is possible that the effort to port a speech recognizer to another language may be reduced with this approach.

2.2. Probabilistic Framework for LBS

The problem of speech recognition can be expressed as the maximization of the posterior probability of sets of phonetic features where each set represents a sound or a phoneme [12]. A set of phonetic features include (1) manner phonetic features represented by landmarks and (2) place or voicing phonetic features found using the landmarks. Mathematically, given an acoustic

Table 3: An illustrative example of the symbols B and L

	/z/	/I/	/r/	/o/	/w/
$U \Rightarrow$	u_1	u_2	u_3	u_4	u_5
	$-sonorant$ $+continuant$ $+strident$ $+voiced$ $+anterior$	$+sonorant$ $+syllabic$ $-back$ $+high$ $+lax$	$+sonorant$ $-syllabic$ $-nasal$ $+rhotic$	$+sonorant$ $+syllabic$ $+back$ $-high$ $+low$	$+sonorant$ $-syllabic$ $-nasal$ $+labial$
$L \Rightarrow$	l_1	l_2	l_3	l_4	l_5
	Fr onset Fr offset	Vowel onset Syllabic peak	SC onset Syllabic dip SC offset	Vowel onset Syllabic peak	SC onset Syllabic dip SC offset

Table 4: APs used in broad class segmentation. f_s : sampling rate, F3 : third formant average, [a,b]: frequency band [aHz,bHz], E[a,b]: energy in the frequency band [aHz,bHz]

Phonetic Feature	APs
Silence	(1) E[0,F3-1000], (2) E[F3,$f_s/2$], (3) ratio of spectral peak in [0,400Hz] to the spectral peak in [400,$f_s/2$], (4) Energy onset [11] (5) Energy offset [11]
$sonorant$	(1) E[0,F3-1000], (2) E[F3,$f_s/2$], (3) Ratio of E[0,F3-1000] to E[F3-1000,$f_s/2$], (4) E[100,400]
$syllabic$	(1) E[640,2800] (2) E[2000,3000] (3) Energy peak in [0,900Hz](4) Location in Hz of peak in [0,900Hz]
$continuant$	(1) Energy onset [11], (2) Energy offset [11], (3) E[0,F3-1000], (4) E[F3-1000,$f_s/2$]

Table 2: Classification results for oral vs nasalized vowels. The classifier was trained on American English (AE) and tested on AE and Hindi.

	9 APs		39 MFCCs	
	AE	Hindi	AE	Hindi
Oral Vowels	68.27	71.38	77.26	88.90
Nasalized Vowels	66.85	56.05	44.68	19.33
Chance Norm. Acc.	67.58	63.72	60.97	54.11

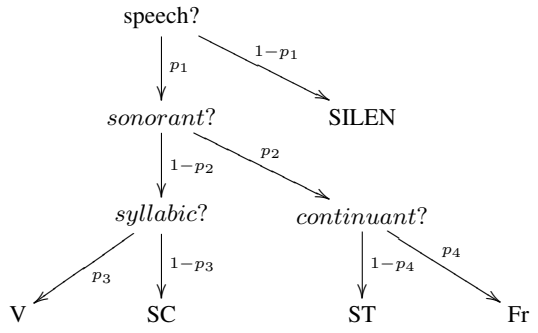

Figure 2: Probabilistic Phonetic Feature Hierarchy

observation sequence O, the problem can be expressed as

$$\hat{U}\hat{L} = \arg \max_{U,L} P(U, L|O) = \arg \max_{U,L} P(L|O)P(U|O, L) \quad (1)$$

where $L = \{l_i\}_{i=1}^{M}$ is a sequence of landmarks and $U = \{u_i\}_{i=1}^{N}$ is the sequence bundles of features corresponding to a phoneme sequence. The meaning of these symbols is illustrated in Table 3 with an example of the digit "zero".

Computation of $P(L|O)$ is the process of probabilistic detection of acoustic landmarks given the acoustic observations and the computation of $P(U|L, O)$ is the process of using the landmarks and acoustic observations to make probabilistic decisions on place and voicing phonetic features. l_i denotes a set of related landmarks that are associated with the same speech sound. For example, the syllabic peak (P) and the vowel onset point (VOP) occur during a vowel. The VOP should occur at the start of the vowel and P should occur during the vowel when the vocal tract is most open.

Table 5: Broad class segmentation results in percent. Correctness (Corr) / Accuracy (Acc) are shown when the system is scored on the basis of numbers of deletions, insertions and substitutions of broad classes. A '-' in a cell means that the particular system was computationally too intensive to get a result from.

	LBS (RBF)	LBS (linear)	HMM
	Corr/Acc	Corr/Acc	Corr/Acc
11 APs	86.2/79.5	84.0/77.1	80.9/73.7
39 MFCCs	-	86.1/78.2	86.8/80.0

To start the speech recognition process, the knowledge-based APs [11] shown in Table 4 for each of the phonetic features - $sonorant$, $syllabic$, $continuant$ - and silence are automatically extracted from each frame of the speech signal. Then, a Linear or Radial Basis Function (RBF) Support Vector Machine (SVM) [13] based binary classifier is applied at each node of the hierarchy shown in Figure 2 (an upper part of the complete hierarchy) such that only the relevant APs for the feature at that node serve as input to the classifier. (Probabilistic hierarchies have been used before in segment-based speech recognition [14].) Probabilistic decisions obtained from the outputs of SVMs are combined with class dependent duration probability densities to obtain one or more segmentations of the speech signal into the broad classes - vowel (V), fricative (Fr), sonorant consonant (SC - including nasals and semi-vowels), stop burst (ST) and silence (SILEN - including stop closures). A segmentation is then used along with the knowledge-based measurements to deterministically find landmarks related to each of the broad class segments. For a fixed vocabulary, segmentation paths can be constrained using broad class pronunciation models to get the probability $P(L|O)$ [12]. APs for the place

Table 6: Broad class results on TIDIGITS (Correct/Accurate in percent)

	LBS (linear)	LBS (RBF)	HMM-MFCC	HMM-AP
Constrained	91.7/82.8	92.6/85.2	92.4/84.3	92.3/85.8
Unconstrained	89.5/64.0	93.0/74.3	88.6/74.1	84.2/72.9

and voicing features are then extracted using the relevant landmarks, and SVMs are applied to get the probabilities of those features. The probabilities of various place and voicing features are combined to get the probability $P(U|OL)$ [12].

The 'si' and 'sx' sentences from the training section of the TIMIT database were used for training and development. For testing, the 'si' and 'sx' sentences from the testing section of the TIMIT database and 2240 isolated digit utterances from the TIDIGITS training corpus were used. The purpose of using the TIDIGITS database in addition to the TIMIT database for testing was to show the performance of LBS in constrained segmentation. The same knowledge-based APs were used to construct a front-end for an HMM-based broad class segmentation system. The results for the TIMIT database are shown in Table 5. The results are also shown for LBS for two different front-ends - AP and MFCC (including MFCCs, their delta and acceleration coefficients which gives a 39 parameter front-end). The performance of all of the systems, except when LBS is used with MFCCs. The results for the TIDIGITS database for both constrained and unconstrained segmentation are shown in Table 6. On moving from unconstrained to constrained segmentation, a similar improvement in performance of the LBS (RBF) and HMM-AP systems can be seen in this table. This result shows that LBS can be constrained by vocabulary in a successful manner similar to the HMM system.

3. Summary

This paper has presented a brief overview of the Landmark-Based approach to speech recognition being followed in our lab. It also lists the advantages that such an approach to speech recognition may have over state-of-the-art approach based on MFCCs and HMMs. The paper has tried to highlight the benefits of both the use of knowledge-based APs instead of MFCCs, and the probabilistic framework for landmark-based recognition instead of HMMs.

The performance of the APs was tested by comparing them with the traditional MFCCs in a digit recognition task using a standard HMM-based system. The results show that the APs are more invariant across databases, speaker and recording condition. Results for a cross-language task have shown that the knowledge-based APs may be more invariant across languages when compared to MFCCs. Next, LBS was compared with an HMM-based system using a broad-class recognition and landmark detection task. The results show that the landmark detector is competitive with a standard HMM-based system in clean, well matched environments. The results also show that LBS can be constrained to recognize only those landmark sequences that are allowed in a limited vocabulary, just like the HMM systems. This makes the integration of LBS with probabilistic pronunciation and language models feasible, allowing the use of LBS in practical recognition tasks. Further, LBS has certain distinct advantages (as have been discussed in the paper) which may make it a very viable alternative to the standard HMM based approach.

4. Acknowledgments

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

5. References

[1] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall (Signal Processing Series), 1993.

[2] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *HTK Documentation*. Microsoft Corporation and Cambridge University Engineering Department, 2006, http://htk.eng.cam.ac.uk/, last viewed by the author on January 30, 2007.

[3] N. Chomsky and N. Halle, *The sound pattern of English*. Harper and Row, 1968.

[4] K. N. Stevens, "Airflow and turbulence noise for fricative and stop consonants," *J. Acoust. Soc. Am.*, vol. 50, pp. 1180–1192, 1971.

[5] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.*, vol. 27, no. 2, pp. 338–352, 1955.

[6] O. Deshmukh, C. Espy-Wilson, and A. Juneja, "Acoustic-phonetic speech parameters for speaker-independent speech recognition," in *Proceedings of ICASSP*, 2002, pp. 593–596.

[7] T. Pruthi and C. Espy-Wilson, "Acoustic parameters for automatic detection of nasal manner," *Speech Communication*, vol. 43, no. 3, pp. 225–239, 2004.

[8] T. Pruthi, "Analysis, vocal-tract modeling and automatic detection of vowel nasalization," Ph.D. dissertation, University of Maryland, College Park, MD, USA, January 2007.

[9] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in *Proceedings of ICASSP*, 1992, pp. 517–520.

[10] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," in *Proceedings of ICSLP*, Banff, Alberta, Canada, 1992.

[11] N. Bitar, "Acoustic analysis and modelling of speech based on phonetic features," Ph.D. dissertation, Boston University, 1997.

[12] A. Juneja, "Speech recognition based on phonetic features and acoustic landmarks," Ph.D. dissertation, University of Maryland, college Park, 2004.

[13] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer Verlag, 1995.

[14] S. Lee, "Probabilistic segmentation for segment-based speech recognition," Master's thesis, Massachusetts Institute of Technology, 1998.