Continuous Optical Automatic Speech Recognition by Lipreading

Alan J. Goldschen*
EECS Department
George Washington University
Washington, DC 20052
ajg@seas.gwu.edu

Oscar N. Garcia†
EECS Department
George Washington University
Washington, DC 20052
garcia@seas.gwu.edu

Eric Petajan
ATT Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974
edp@research.att.com

Abstract

We describe a continuous optical automatic speech recognizer (OASR) that uses optical information from the oral-cavity shadow of a speaker. The system achieves a 25.3 percent recognition on sentences having a perplexity of 150 without using any syntactic, semantic, acoustic, or contextual guides. We introduce 13, mostly dynamic, oral-cavity features used for optical recognition, present phones that appear optically similar (visemes) for our speaker, and present the recognition results for our Hidden Markov Models (HMMs) using visemes, trisemes, and generalized trisemes. We conclude that future research is warranted for optical recognition, especially when combined with other input modalities.

1 Introduction

This paper investigates a method of performing continuous automatic speech recognition using only optical information obtained from the shadow of the oral-cavity region of a speaker. The system attempts to determine the correct spoken sentence using images of the oral-cavity captured by a camera.

Our research approach parallels the building of continuous acoustic automatic speech recognizers [12], [20]. The important features of the oral-cavity region of our speaker for optical recognition are first determined. These features are vector quantized into a codebook using a clustering algorithm. The HMMs use the sequence of codebook entries to capture optical knowledge about the structure of the spoken sentence.

*Employed with Martin Marietta Laboratories in Baltimore, MD

†Presently on leave as Program Director of the Interactive Systems Program at the National Science Foundation

2 Description of the system

The optical recognizer of the overall system consists of two components. The first component of the optical recognizer is the optical processor that converts the optical features of the oral-cavity region into a sequence of codewords. The second component of the optical recognizer is the (HMM) linguistic decoder that converts the sequence of codewords into a recognized sentence, with each sentence identified as a string of visemes.[1]

Our database consists of optical recordings captured to disk at 60 frames-per-second of one speaker with a beard and mustache reading 450 TIMIT sentences [14]. The recording of the speaker's lower facial region was obtained with a head-mounted harness containing a camera, two side-mounted incandescent lights with a microphone. This very large database contains each sentence in ASCII, acoustical, and optical formats. A significant amount of image processing has been done to this database [16] [17] [18]. Figure 1 illustrates a sequence of the optical database frames. Before calculating the most significant features from an image frame, some noise is removed from the image frame. The oral-cavity region is rotated in an image frame by the angle that the horizontal axis produces when connecting an imaginary line between the two nostrils. This angular rotation ensures that the oral-cavity region is horizontal. The oral-cavity is then centered within the image frame. Next, noise pixels are removed from the image using a median filter that iterates over an image until no further filtering occurs.

Only 67 of the 450 TIMIT sentences spoken by our speaker were phonically transcribed with time marks (hand-segmented); for the remaining 383 sentences, we assumed that the phonetic transcriptions of the

[1] A viseme is defined in [6] as a particular sequence of oral-cavity movements (shapes) that corresponds to a phoneme.

1058-6393/95 $4.00 © 1995 IEEE

TIMIT database[14] were valid. 150 sentences were randomly chosen to test the system and the remaining 300 sentences were used to train the system.[2] The 300 training sentences, include the 67 hand-segmented sentences, were used to build context-independent viseme HMMs, context-dependent triseme[3] HMMs, and generalized triseme[4] HMMs.

3 Optical processor

The optical processor of the optical recognizer receives a sequence of optical images similar to the contour or edge sequences depicted in Figure 1, and converts these optical images into a sequence of codewords.

3.1 Extracting oral-cavity features

The first step in designing the optical processor is to determine those features from the oral-cavity region that contribute the most significant information to OASR.

The seven (static) features listed in Table 1 are calculated from the oral-cavity region. In addition to computing these seven (static) features, we investigate the dynamics of each feature. We first calculate the change of each feature between successive frames (first derivative) and then calculate the corresponding change of change (second derivative). We also consider the magnitude of the first and second derivatives because we want to know if the direction of change is an important oral-cavity feature. Hence, the system computes 35 initial features from each oral-cavity image in the 67 hand-segmented training sentences.

3.2 Analyzing oral-cavity features

We then analyze and try to reduce the number of oral-cavity features. For this process we create a correlation matrix and perform a principal component analysis using all of the images in the 67 hand-segmented sentences (for a total of 11214 images). Using our correlation matrix and feature selection algorithms in [10], we decided to retain the following 13 features from our original 35 features: *I-Blob, Rounding, Width, Area$_I$, Height$_I$, Rounding$_I$, Area$_{II}$, Height$_{II}$, IBlob$_{II}$, Perimeter$_{II}$, |Rounding$_I$|, |Height$_{II}$|, |Perimeter$_{II}$|*. Although we kept three static oral-cavity features, notice that most of the features kept pertain to the derivatives, confirming our belief that the dynamics of the oral-cavity features are important for computerized lipreading analysis [7] [8].

[2]Note that we randomly selected our training sentences thereby admitting that we are not taking advantage of the phonically balanced (first and higher statistical) properties associated with the TIMIT database. We chose to randomly select the training and testing sentences to avoid any personal biases, recognizing that we may have paid a price in the results obtained by not using the balanced distribution in training and testing.

[3]A triseme is a triplet of visemes that occurs in our database.

[4]A generalized triseme is obtained from similar context-dependent triseme HMMs that have been clustered together.

3.3 The codebook

Each image frame of a sentence is represented as a vector containing the values for the 13 oral-cavity features kept. The next step is to build a codebook to minimize the number of possible vectors since some of the 13-dimensional feature vectors are in close proximity to each other. Each vector is to be associated (clustered) with a codeword. These codewords are the output symbols, or observation sequences, for training and testing our HMMs. We used a K-Means non-hierarchical clustering algorithm [1] [8] to generate a codebook of 64 codewords using the Euclidean distance as our metric. A codebook size of 64 codewords was chosen as a compromise between the representation of the vector space and having enough data to train each HMM observation symbol table. Thus, each of the 450 optical sentences becomes a sequence of codebook entries.

4 The linguistic decoder

The linguistic decoder of the optical processor accepts a sequence of codebook entries from the optical processor and uses a-priori knowledge about the grammar from the set of test sentences.

4.1 Determining visemes

To determine the phones that correspond to a viseme for our speaker, we train HMMs using the Forward-Backward (F-B) algorithm [20] [13] on each phone of the 67 hand-segmented sentences from the optical database.

We first divide our phones into three groups to facilitate the clustering in a more homogeneous space. The first group contains consonant phones where the closures are modeled with their respective consonant phones. The second group contains the consonant phones with the closures modeled separately, and the third group contains the vowel and diphthong phones.

The clustering process uses the Average Linkage hierarchical clustering algorithm[1][5] and the HMM similarity metric D (Equation 1) as defined in [11].

$$D(\Omega_1, \Omega_2) = \frac{1}{T} * [\log \Pr(O^2|\Omega_1) - \log \Pr(O^2|\Omega_2)] \quad (1)$$

Equation 1 indicates how likely observation sequence O^2 (used to train HMM Ω_2) compares to another HMM Ω_1, and is the ratio of $\Pr(O^2|\Omega_1)$ to $\Pr(O^2|\Omega_2)$. The conditional probabilities are computed using the HMM forward algorithm [20] [13]. Because D of equation 1 is non-symmetric, we use Equation 2 to compute the average similarity D_s between HMMs.

$$D_s(\Omega_1, \Omega_2) = \frac{1}{2} * [D(\Omega_1, \Omega_2) + D(\Omega_2, \Omega_1)] \quad (2)$$

Table 2 yields the results from the clustering algorithm. It is important to note that most of the consonant viseme groups depicted in Table 2 were obtained from the experiences of expert human lipreaders or lipreading researchers such as [2], [9], [15], [21], and [23]. Unlike other researchers and like our research, Finn [5] uses an algorithmic objective approach to determine her consonant visemes.

The viseme groups that emerge when the closures are modeled with their respective consonants are depicted in the *Goldschen* column of Table 2. These viseme groups appear to be fairly consistent with the results of other research. We are encouraged to find that using our computer algorithms and data we obtain results similar to those contained in the other studies of human lipreaders.

Despite the previous results, our data includes the segmentation of closures that are separated from their consonant phones. Using this data, the same viseme groups emerge as before except that the system obtains the viseme group *(/bcl/,/m/,/pcl/)* instead of *(/b/,/m/,/p/)* and obtains *(/b/,/p/,/r/)* instead of *(/r/)*. These results clearly indicate (as we suspected, although other researchers had not considered them) that closure affects the grouping of visemes.

The left column of Table 4 depicts the viseme grouping of our vowel phones. In most cases the solitary vowel or diphthong phone forms its own viseme group with the exception of the viseme groups *(/ax/,/ih/,/iy/)* and *(/ae/,/eh/)*. Recalling the vowel triangle [19],the viseme group *(/ax/,/ih/,/iy/)* includes vowels associated with similar tongue height (high). It appears, at least for our speaker, that the high back vowel */ax/* is not distinguishable from the high front vowels */iy/* and */ih/*. The vowel triangle also shows that */ae/* and */eh/* are vowels with similar tongue height (low).

The phone-to-viseme mappings for our speaker are summarized in Table 4. We decided to separate the phones */r/* and */w/* into two groups based on our phonetic knowledge [8].

[5]The Average Linkage clustering algorithm is similar to the Complete Linkage and the Single Linkage clustering algorithms.

4.2 Summary of HMM training

Figure 2 summarizes the overall training process. The 67 hand-segmented sentences are used to train the initial 56 phone HMMs using the F-B algorithm. Similar phone HMMs are clustered into 35 viseme HMMs using the Average Linkage clustering algorithm. The 67 phone segmented sentences are again used with the F-B algorithm to create the context-independent viseme HMMs. The 300 training sentences are used to generate the 2702 context-dependent triseme HMMs using the F-B algorithm with embedded re-estimation.

We built generalized trisemes to reduce the number of context-dependent triseme HMMs, thereby having more training samples per HMM. Since perfect phone boundaries did not exist for all of the trisemes, the trisemes could not be clustered as done previously. Instead the 2702 context-dependent HMMs are used to generate the observation sequences required for our HMM similarity metric of Equation 2. This is a novel approach to solving the problem. After generating the observation sequences for each triseme, we cluster only those trisemes that share the same middle viseme and include the middle viseme in this group. The Average Linkage clustering algorithm merges similar trisemes that have the same middle viseme. Next, the 300 training sentences and the F-B algorithm with embedded re-estimation are used to generate the 934 generalized triseme HMMs.

5 Testing the OASR system

After creating the optical recognizer (consisting of both the optical processor and the linguistic decoder), we test the performance using 150 test sentences.

From the set of 150 test sentences in our optical database, a sentence to be recognized is chosen and the appropriate codebook entries are determined from the optical features. The recognized sentence is the sentence that has the maximum Viterbi probability [20] [13] from all our 150 sentences. This process repeats for each of the 150 test sentences. Table 3 yields

the percentage of correctly recognized test sentences using the context-independent viseme HMMs, the context dependent triseme HMMs, and the generalized triseme HMMs.

It is difficult to judge the quality of our results because no other system performs continuous, automatic speech recognition using optical information. [3], [4], [22] mention that audiologists have performed isolated-word speech recognition studies in an attempt to determine the percentage of correctly recognized words. When using only acoustic information, individuals who possess normal hearing will correctly recognize about 40 percent of the isolated nonsense words. Expert lipreaders, using only visual information, will correctly recognition about 30 percent of these nonsense words. Yet, using both acoustical and optical information, individuals who possess normal hearing and are expert lipreaders, will correctly recognize about 70 percent of these nonsense isolated words [3], [4], [22]. In contrast to the 30 percent recognition rate for nonsense words by expert lipreaders, our system achieved a 25 percent recognition rate for a group of sentences without the aid of acoustic and grammatical information.

6 Conclusions

Using HMM and hierarchical clustering algorithms, our system identifies viseme groups that are consistent with the consonant viseme groups selected by expert human lipreaders and the vowel triangle [19]. Our result, however, indicates that when using closures, the mapping of phones to visemes changes slightly. Future researchers in automatic speech recognition may want to use the vowel visemes to augment the acoustic information because the vowels appear to be characterized by oral-cavity features. This result may be useful for the teaching of lipreading because closures have often been ignored.

We present dynamic features, some never introduced before, to describe the actions of the oral-cavity region during speech. We use a correlation matrix and a principal component analysis to reduce the oral-cavity feature space from 35 to 13 features. After finding the oral-cavity region from an image frame, most of the 13 features can easily be calculated with modern machines. We believe, additionally, that these dynamic features are very important for lipreading research, especially for human lipreading. Most students are taught to lipread by the placement of (static) oral-cavity features. Our research indicates, however, that the movement of (dynamic) oral-cavity features allow for successful lipreading.

We expect future work to examine whether additional oral-cavity features could be used, and focus on utilizing a trigram, bigram or word-pair grammar. We recommend that speaker-independent OASR be investigated since most successful continuous acoustic automatic speech recognition systems using HMMs work quite well with multiple speakers. An investigation with multiple speakers would greatly contribute to this research by confirming the consistency of the viseme groups obtained herein. Finally, we suggest that this system augment an existing or future acoustic automatic speech recognition system, especially in environments when the acoustic signal is noisy or degraded.

References

[1] Michael Andeberg. *Cluster Analysis for Applications.* Academic Press, New York, NY, 1973.

[2] J. Burchett. *Lipreading: A Handbook of Visible Speech.* The Royal National Institute for the Deaf, London, England, 1965.

[3] Roberta Cerio. Personal communications, February 1989.

[4] Orin Cornett. Personal communications, February 1989.

[5] Kathleen Finn. *An Investigation of Visible Lip Information to be used in Automatic Speech Recognition.* PhD thesis, Georgetown University, 1986.

[6] Cletus Fisher. Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11:796–804, 1968.

[7] Oscar Garcia, Alan Goldschen, and Eric Petajan. Feature extraction for optical automatic speech recognition or automatic lipreading. Technical Report IIST-92-32, The George Washington University, November 1992.

[8] Alan Goldschen. *Continuous Automatic Speech Recognition by Lipreading.* PhD thesis, The George Washington University, Washington, D-C, 1993.

[9] Elizabeth Hazard. *Lipreading: For the Oral Deaf and Hard-of-Hearing Person.* Charles C. Thomas, Springfield, Illinois, 1971.

[10] I.T. Jolliffe. *Principal Component Analysis.* Springer-Verlag, New York, NY, 1986.

[11] Biing Juang and Lawrence Rabiner. A probabilistic distance measure for hidden markov models. *ATT Technical Journal*, 64(2):391–408, February 1985.

[12] Kai Fu Lee. *Automatic Speech Recognition: The Development of the Sphinx System.* PhD thesis, Carnegie-Mellon University, Pittsburgh, PA 15213, 1989.

[13] Stephen Levinson, Lawrence Rabiner, and Man Mohan Sondhi. An introduction to the application of theory of probabilistic function of a markov process to automatic speech recognition. *The Bell System Technical Journal*, 62(4):1035–1074, April 1983.

[14] NIST, Gaithersburg, MD 20899. *DARPA Timit CD-ROM*, November 1988.

[15] Elizabeth Nitchie. *New Lessons in Lipreading.* J.B. Lippincutt Company, New York, NY, 1950.

[16] Eric Petajan. *Automatic Lipreading to Enhance Speech Recognition.* PhD thesis, University of Illinois at Urbana-Champaign, 1984.

[17] Eric Petajan. Automatic lipreading to enhance speech recognition. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 40–47, San Francisco, CA, 1985. IEEE.

[18] Eric Petajan, Bradford Bischoff, David Bodoff, and N. Michael Brooke. An improved automatic lipreading system to enhance speech recognition. In *CHI-88*, pages 19–25. ACM, 1988.

[19] Gordan Peterson and Harold Barney. Control methods used in a study of the vowels. *Journal of Acoustical Society of American*, 24:175–184, March 1952.

[20] Lawrence Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In Alex Waibel and Kai-Fu Lee, editors, *Readings in Speech Recognition*, pages 267–296. Morgan Kaufmann Publishers, Inc., 1990.

[21] Quentin Summerfield. Some preliminaries to a comprehensive account of audio-visual speech perception. In Barbara Dodd and Ruth Campbell, editors, *Hearing by Eye: The Psychology of Lipreading*, pages 3–51. Lawrence Earlbaum Associates, 1987.

[22] Henry Tobin. Personal communications, February 1989.

[23] Brian Walden, Robert Prosek, Allen Montgomery, Charlene Scherr, and Carla Jones. Effects of training on the visual recognition of consonant. *Journal of Speech and Hearing Research*, 20:130–145, 1977.

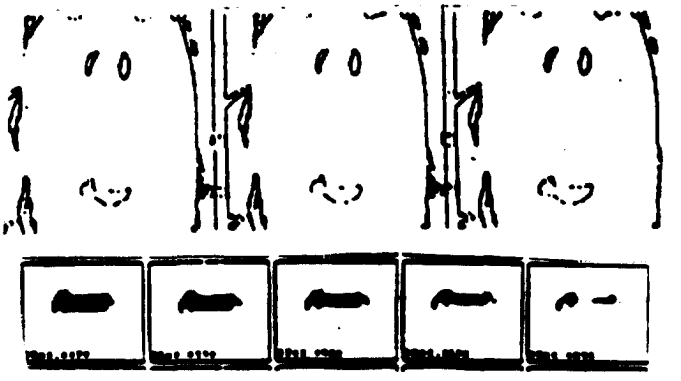

Figure 1: Sample Contour (Edge) and Image (Oral-Cavity) Frames: The top part of the figure consists of contours or edges depicting the nostrils, oral-cavity, and facial regions. The bottom part of the figure depicts a sequence of oral-cavity region frames as binary images.

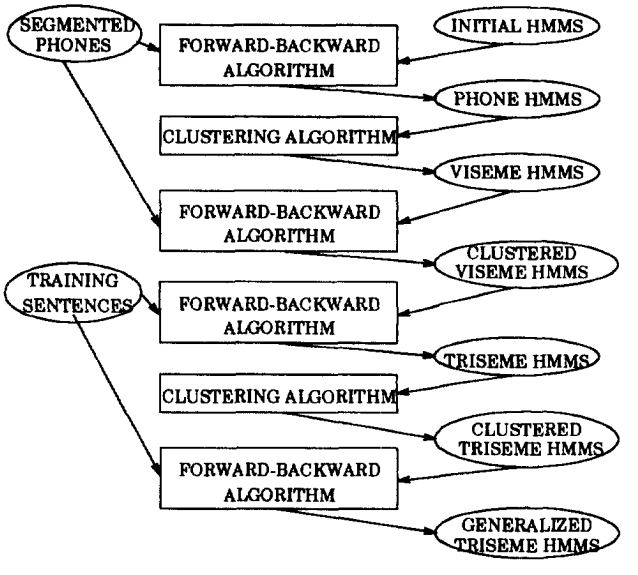

Figure 2: Summary of the Training Process.

Area	the number of black pixels in an image frame
CBlob	(contour blob) the number of connected (closed) regions in a contour frame
Height	the vertical distance between the maximum y and minimum y ordinates of the shadow of the oral-cavity
IBlob	(image blob) the number of regions in a binary image frame (after filtering)
Perimeter	the number of pixels along the edge of the shadow of the oral-cavity
Rounding	the radio of the width to the height
Width	the horizontal distance between the maximum x and minimum x abscissa of the shadow of the oral-cavity

Table 1: Initial Static Oral-Cavity Features Calculated From Each Image Frame.

HMMs	Number Correct	Percent Correct
Context-Independent Viseme HMMs	3	2
Context-Dependent Triseme HMMs	19	12.7
Generalized Triseme HMMs	38	25.3

Table 3: Number and Percentage of Correctly Recognized Sentences.

Burchett	Hazard	Nitchie	Summerfield	Walden	Finn	Goldschen
b, m, p	b, m, p	b, m, p	b, m, p	b, m, p	b, m, p	b, m, p
r			r	r	r	r
f, v	f, v	f, v	f, v	f, v	f, v	f, v, w
w	w		w	w	w	w
th		th	th	th, dh	th	th
d, n, t	d, n, t	d, n, t	d, t	d, g, j, k, n, t	d, n, l, t	d, dh, epi g, k, l n, t
l		l	l	l		
k, g	k, g	k, g	n, k, g		k, g	
s, z	s, z	s, z	s, z	s, z	s, z	s, sh, z
ch, j, sh, zh	ch, j, sh	ch, jh, zh	sh, zh	sh, zh	ch, jh, sh, zh	zh
		y	y		y	y
h		h			hh	hh
						hv
						ch
						dx, nx, q
						en
						jh
						ng
						h#

Table 2: Mapping of Consonant Phones to Visemes: This table assumes that the closure phones are included with their respective consonants.

Vowels and Diphthongs	Consonants
aa	b, p
ae, eh	bcl, m, pcl
ah	ch
ao	d, dcl, g, gcl, k, kcl, l, n, t, tcl
aw	dh, epi
ax, ih, iy	dx, nx, q
axr	f, v
ay	en
er	hh
ey	hv
ix	jh
ow	ng
oy	r
uh	s, sh, z
uw	th
ux	w
	y
	zh
	h#

Table 4: Final Phone To Viseme Mapping.