A Multi-Stream Approach to Audiovisual Automatic Speech Recognition

(Invited Paper)

Mark Hasegawa-Johnson
University of Illinois at Urbana-Champaign,
jhasegaw@uiuc.edu

***Abstract*— This paper proposes a multi-stream approach to automatic audiovisual speech recognition, based in part on Hickok and Poeppel's dual-stream model of human speech processing. The dual-stream model proposes that semantic networks may be accessed by at least three parallel neural streams: at least two ventral streams that map directly from acoustics to words (with different time scales), and at least one dorsal stream that maps from acoustics to articulation. Our implementation represents each of these streams by a dynamic Bayesian network; disagreements between the three streams are resolved using a voting scheme. The proposed algorithm was tested using the CUAVE audiovisual speech corpus. Results indicate that the ventral stream model tends to make fewer mistakes in the labeling of vowels, while the dorsal stream model tends to make fewer mistakes in the labeling of consonants; the recognizer voting scheme takes advantage of these differences to reduce overall word error rate.**

I. Introduction

This paper describes ongoing experiments using an articulatory feature model (AFM) of automatic audiovisual speech recognition (AVSR). The model itself, and its word accuracy, were first described in in [1], [2]; the model and its word accuracy are reviewed below in Sec. III and Fig. 2. This paper offers a new analysis of Fig. 2, based on the dual-stream human speech processing model of Hickok and Poeppel [3]. Hickok and Poeppel have proposed that the cortical conceptual network may be accessed in parallel by at least three different neural streams: a right ventral stream that performs lexical access from acoustic phonological representations, possibly on a syllabic time scale; a left ventral stream that performs lexical access from acoustic phonological representations, possibly on a segmental time scale; and a left dorsal stream that touches on articulation. Our AVSR strategy is based on the parallel combination of three different recognition models, which we propose to be loosely comparable to the analyses of the right ventral, left ventral, and left dorsal neural streams.

II. The Dual-Stream Model

Despite a century of research, there is still much disagreement about the functions of cortical structures supporting speech processing. Hickok and Poeppel [3] propose that the disagreement stems from a common false assumption. Many neuropsychological studies assume that one may learn about human speech recognition (the mapping from sound to meaning) by studying human speech perception (the mapping from sound to articulatory or phonological units). Indeed, Hickok and Poeppel argue, it is clear that one *may* associate concepts with articulatory gestures, otherwise it would be impossible to learn new words. The false assumption, they argue, is that one *must* associate concepts with articulatory gestures. In clinical studies, recognition and perception doubly dissociate: there are patients who are able to recognize words but not nonsense syllables, and vice versa. Based on the double dissociation between speech recognition and speech perception, and on a review of other neuropsychological results, Hickok and Poeppel propose a dual-stream model of human speech processing (Fig. 1).

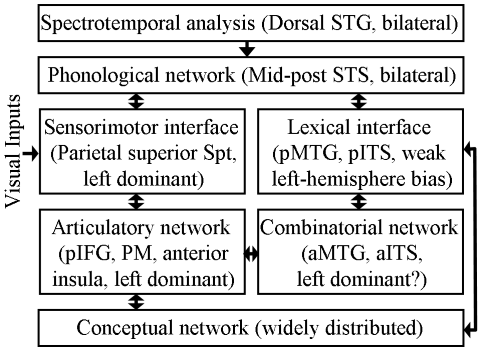
Spectrotemporal analysis (Dorsal STG, bilateral)
Phonological network (Mid-post STS, bilateral)
Visual Inputs
Sensorimotor interface (Parietal superior Spt, left dominant)
Lexical interface (pMTG, pITS, weak left-hemisphere bias)
Articulatory network (pIFG, PM, anterior insula, left dominant)
Combinatorial network (aMTG, aITS, left dominant?)
Conceptual network (widely distributed)

Fig. 1. Dual-stream model of human speech processing, after [3] Fig. 1. The dorsal stream (left column) "maps sensory or phonological representations onto articulatory motor representations;" the ventral stream (right column) "maps sensory or phonological representations onto lexical conceptual representations." STG, superior temporal gyrus; STS, superior temporal sulcus; MTG, middle temporal gyrus (aMTG, anterior; pMTG, posterior); ITS, inferior temporal sulcus (aITS, anterior; pITS, posterior); Spt, Sylvian fissure at parieto-temporal boundary; pIFG, posterior inferior frontal gyrus; PM, premotor cortex.

The dual-stream model proposes that human speech processing occurs in two or more parallel neural streams. Both neural streams are activated by phonological representations in the superior temporal sulcus (STS), and these in turn are activated by spectrotemporal representations in the superior temporal gyrus (STG). One of the two streams is ventral, involving structures in the middle temporal gyrus (MTG) and inferior temporal sulcus (ITS) bilaterally, and "maps sensory or phonological representations onto lexical conceptual representations" [3]. Damage to the posterior temporal lobe has consistently been demonstrated to cause auditory comprehension deficits, implying that pMTG and pITS are important in the mapping from phonology to meaning. Anterior structures are more active when subjects listen to sentences as compared to unstructured lists of words or sounds, implying that aMTG and aITS support syntax and/or combinatorial semantics. Concept relationships are robust to temporal lobe trauma; [3] proposes that conceptual representations are distributed widely throughout the cortex.

The second stream is dorsal, involving structures in the parietal and frontal lobes, and serves as the primary stream by which "sensory or phonological representations" activate articulatory units in the pre-motor cortex (PM) and posterior inferior frontal gyrus (pIFG, e.g., Broca's area). [3] proposes, in particular, that circuits in the Sylvian fissure at the boundary of parietal and temporal cortex (Spt) capture input from multiple sensory modalities, and organize that input for the activation of vocal tract regions in PM and pIFG. Modalities processed by Spt include speech, visual text, music, and humming. Region Spt is part of the planum temporale (PT); other regions of PT are activated by sensory inputs including visible speech, spatial audio, sign language, and visible motion.

III. Audiovisual Speech Recognition

This section describes an automatic audiovisual speech recognition (AVSR) system that matches many features of the dual-stream model, as well as matching available information about the human processing of visible speech. Hickok and Poeppel do not mention visible speech, other than to note that it activates the planum temporale, but many other authors have studied audiovisual integration. The McGurk effect [4] demonstrates that pre-conscious processes mediate between conflicting auditory and visual cues in order to find a percept (a non-word or word) that best matches both audio and visual inputs. Silent lipreading activates the primary auditory cortex [5], therefore it is possible that visible speech may induce activation in all of the cortical structures named in Fig. 1, exactly as they would be activated by audible speech. Activation of primary auditory cortex by visible speech, and the McGurk effect, taken together, imply that visible and audible speech information must be integrated pre-phonologically—either these sources of information are integrated in the primary auditory cortex, or they are integrated by the acoustic-to-phonological mappings computed (according to [3]) by the STS. Evidence from audiovisual asynchrony supports the latter explanation. Errors in synchronization between the audio and video playback heads of a videotape are not detectable, by most subjects, at time scales shorter than the duration of a typical phoneme (below about 60ms); if asynchrony between the audio and video playback exceeds the 60ms threshold, most subjects find it both noticeable and extremely annoying [6].

In the dual-stream model, "portions of the STS are important for representing and/or processing phonological information." The nature of the STS representation is not specified, but may include "distinctive features, segments (phonemes), syllabic structure, phonological word forms, grammatical features and semantic information" [3]. In our algorithm, our model of the STS computes the likelihoods of three different phonological units: phoneStates, visemeStates, and audiovisual gesture vectors. The audio signal is represented by a 39-dimensional vector $\vec{x}_t$ computed once per 10ms, containing MFCCs, energy, deltas and delta-deltas. The video signal is represented by a 70-dimensional vector $\vec{y}_t$ interpolated to a 10ms sample period, containing the first 35 discrete cosine transform coefficients of the lip image, and their deltas. A phoneState, q_t, is one third of a phone (there are three phoneStates per TIMIT phone), and a visemeState, v_t, is a third of a viseme; their relationship to the audio and video features is given by:

$$p(\vec{x}_t|q_t) \quad \propto \quad \left(\sum_{k=1}^{K} c_{qk}\mathcal{N}(\vec{x}_t;\mu_{qk},\Sigma_{qk})\right)^{1-\gamma}, \qquad (1)$$

$$p(\vec{y}_t|v_t) \quad \propto \quad \left(\sum_{k=1}^{K} c_{vk}\mathcal{N}(\vec{y}_t;\mu_{vk},\Sigma_{vk})\right)^{\gamma}, \qquad (2)$$

where $\mathcal{N}(x;\mu,\Sigma)$ is the normal density with mean μ and diagonal covariance matrix Σ, c_{qk} and c_{vk} are mixture weights, and γ is a video stream weight chosen to minimize word error rate.

A "gesture vector" $\vec{r}_t = [l_t, u_t, g_t]'$ specifies the articulatory gestures currently being implemented by the lips (l_t), by the tongue (u_t), and by the glottis and velum (g_t). The list of gestures that may be implemented by each articulator is adapted from [7]; an exact list is given in [2]. The relationship between a gesture vector and the audiovisual signal is given by

$$p(\vec{x}_t,\vec{y}_t|\vec{r}_t) = p(\vec{x}_t|\vec{r}_t)^{1-\gamma}p(\vec{y}_t|\vec{r}_t)^{\gamma} \qquad (3)$$

where $p(\vec{x}_t|\vec{r}_t)$ and $p(\vec{y}_t|\vec{r}_t)$ are each mixture Gaussian PDFs of the form given in Eqs. 1 and 2.

The ventral stream in [3], and areas pMTG and pITS in particular, "are involved in the mapping between phonological representations in the STS and widely distributed semantic representations" [3]. In our implementation, the mapping to any given word string ($W = [w_1,\ldots,w_T]$) given a corresponding phoneState string ($Q = [q_1,\ldots,q_T]$) and visemeState string ($V = [v_1,\ldots,v_T]$) is computed using a coupled hidden Markov model (CHMM [8]):

$$p(Q,V|W) = \prod_{t=1}^{T} p(q_t|q_{t-1},v_{t-1},w_t)p(v_t|q_{t-1},v_{t-1},w_t) \qquad (4)$$

The transition probabilities $p(q_t|q_{t-1},v_{t-1},w_t)$ and $p(v_t|q_{t-1},v_{t-1},w_t)$ are learned from training data, subject to a designer-specified limit on the inter-modal asynchrony: $|q_t - v_t| \le \delta_{max}$.

The dual-stream model of speech processing proposes that there is "at least one pathway in each hemisphere that can process speech sounds sufficiently well to access the mental lexicon," and that there may be differences between the time scales of lexical integration in the two hemispheres: "neural mechanisms for integrating information over longer timescales are predominantly located in the right hemisphere, whereas mechanisms for integrating over shorter timescales might be represented more bilaterally" [3]. In our implementation, bilateral lexical access is modeled by the use of two CHMM speech recognizers in parallel. The longer time scale of the right hemisphere is modeled by the use of a larger δ_{max} in the right-hemisphere CHMM: in our current implementation, $\delta_{max} = 2$ in the right ventral stream, and $\delta_{max} = 1$ in the left ventral stream.

The dorsal stream in [3], and area Spt in particular, "is involved in translation between sensory codes and the motor system," and "this sensory representation can then be used to guide motor articulatory sequences" [3]. Hickok and Poeppel specifically remove area Spt from the process of speech recognition (which they attribute to the ventral stream), but we propose that feedback from articulatory representations to the lexicon may be useful in speech recognition, for the following reasons. In casual speech, talkers rarely produce a word using exactly the phonemes specified in the dictionary. In 3.5 hours of speech phonetically transcribed by the Switchboard Transcription Project, among words spoken at least five times, the mean number of distinct pronunciations per word is 8.8; the dictionary pronunciation is rarely the most common pronunciation, and for many words, the dictionary pronunciation is never uttered [7]. The theory of Articulatory Phonology [9] gives a comprehensive account of pronunciation variability, according to which the fundamental units of phonology are articulatory gestures (e.g., "lips open," "tongue tip close"). The mapping from acoustics to gesture combinations may be computed in the temporal lobe, as suggested in Eq. 3, but gesture sequence probabilities remain arbitrary unless they are grounded in feedback from the articulatory motor system. It is proposed, therefore, that there is a circuit somewhere in the connection between the temporal lobe and the dorsal speech stream that computes the probability of a gesture sequence ($R = [\vec{r}_1, \ldots, \vec{r}_T]$) given any particular hypothesized word string (W). In our implementation, this computation is represented using a three-chain CHMM:

$$p(R|W) = \prod_{t=1}^{T} p(l_t|\vec{r}_{t-1}, w_t)p(u_t|\vec{r}_{t-1}, w_t)p(g_t|\vec{r}_{t-1}, w_t) \tag{5}$$

The dual-stream model "proposes that there are multiple routes to lexical access, which are implemented as parallel channels" [3]. Parallel automatic speech recognizers, each performing a different type of acoustic analysis, have been combined recently using algorithms such as ROVER [10]. ROVER aligns the transcription outputs of multiple recognizers, in order to form a word lattice; ambiguities in the lattice (alternate explanations for the same period of time) are resolved by majority voting. Voting and weighted voting are easily implemented in a neural or neuronal network (see, e.g., [11]), therefore we consider it reasonable to propose the ROVER algorithm as a model for the resolution of differences among lexical access results of the left ventral, right ventral, and left dorsal speech processing streams.

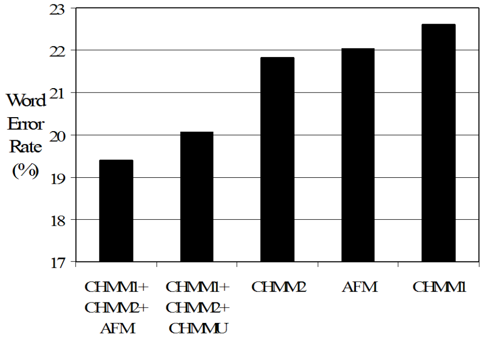

Fig. 2. Word error rates of three original systems (audiovisual CHMM systems with $\delta_{max} = 1$ and $\delta_{max} = 2$, and articulatory feature model (AFM) with $\delta_{max} = 2$) and two ROVER system combinations, averaged across all SNRs (after [2]).

IV. Experimental Results

The three AVSR subsystems described in Sec. III were implemented in the graphical modeling toolkit (GMTK), and tested using connected digits (digits with silence between) from the CUAVE database [12]. Acoustic noise was added at signal to noise ratios (SNRs) of ∞, 12, 10, 6, 4, and -4dB prior to the computation of MFCCs. System combination was implemented using the NIST ROVER toolkit [10]. Systems were trained using noise-free data from 13 talkers. Video weights were determined in order to minimize word error rate on data from 6 talkers (10 digits$\times$6 talkers$\times$5 tokens per digit=300 word tokens per SNR), and the resulting development-test word error rates (averaged across all six SNRs) are given in Fig. 2.

The leftmost bar in Fig. 2 shows the word error rate achieved by a system combination using two audiovisual CHMM systems ($\delta_{max} = 1$ and $\delta_{max} = 2$) and one articulatory-feature model (AFM: a three-chain CHMM with articulatory state space vector, as specified in Eqs. 3 and 5). The second bar shows the result of system combination in which the AFM has been replaced by another audiovisual CHMM ($\delta_{max} = \infty$). The word error rates achieved using system combination are lower than those achieved without system combination (MAPSSWE test [13], $p < 0.05$). There is no significant difference between the two ROVER system combinations, but, as shown in the figure, system combination including the AFM tends to yield lower WER than system combination without the AFM.

Table I lists the most common substitution errors made by the best audiovisual CHMM and the best AFM. The most striking feature of the two lists is their similarity, but there are a few differences. Differences between the columns suggest that the audiovisual CHMM is more likely to make substitutions that maintain the same vowel, while the AFM is more likely to make substitutions that maintain features of the consonants. For example, the second most common error made by the audiovisual CHMM is the substitution of "nine" in place of "five," maintaining the vowel; the second most common error

TABLE I

MOST COMMON SUBSTITUTION ERRORS MADE BY THE AUDIOVISUAL CHMM (LEFT COLUMN) AND THE ARTICULATORY FEATURE MODEL (RIGHT COLUMN). NUMBER OF OCCURRENCES IN PARENTHESES; TEST CORPUS CONTAINED 1800 WORD TOKENS.

CHMM	AFM
one→nine (56)	one→nine (52)
five→nine (32)	three→nine (31)
four→nine (27)	four→nine (27)
three→nine (26)	five→nine (19)
four→one (17)	two→six (18)
seven→six (16)	four→one (14)
two→nine (16)	two→nine (14)

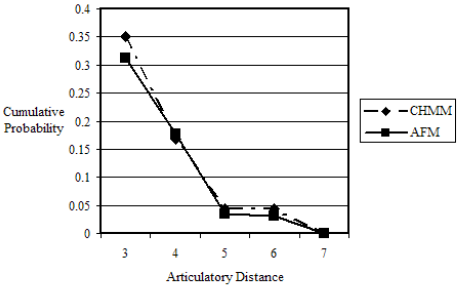

Fig. 3. Cumulative probability of a substitution error with consonantal articulatory distance greater than or equal to the number shown.

made by the AFM is the substitution of "nine" in place of "three," maintaining the tongue as primary articulator of the initial consonant. The fifth most common error made by the audiovisual CHMM is the substitution of "one" for "four," maintaining the vowel; the fifth most common error made by the AFM is the substitution of "six" for "two," maintaining the tongue as primary articulator of the initial consonant.

Fig. 3 shows the cumulative probability that a substitution error by either algorithm exceeds any given consonantal articulatory distance. Articulatory distance is computed as the number of consonants with altered primary articulatory (lips vs. tongue), plus the number of consonants with altered voicing (voiced vs. unvoiced). Each word has up to three consonants, thus the maximum possible articulatory distance between the uttered and recognized words is 6 features. There is a small, marginally significant difference ($t_{1799} = 1.15$, $p < 0.15$) between the average articulatory distance of AFM errors (2.28) and the average articulatory difference of CHMM errors (2.32).

V. CONCLUSIONS

This paper makes three specific proposals. First, it is proposed that an automatic speech recognition system should make use of multiple independent lexical access algorithms, as in the dual-path model of [3], and that disagreements between the different lexical access algorithms may be resolved by voting. Second, we propose that at least one of the lexical access algorithms should impose articulatory plausibility constraints on the recognized transcription; in our implementation, articulatory plausibility constraints are imposed by a dynamic Bayesian network AFM. Third, we propose that acoustic and visual observations should be integrated at the level of the acoustic-to-phonological mapping, either by the use of independent phoneme and viseme hidden states, or by the use of a multi-stream observation PDF dependent on the current values of the hidden articulatory features. The AFM and the audiovisual CHMM learn slightly different facts about the training corpus; the ROVER algorithm is able to take advantage of the differences between the AFM and CHMM in order to reduce total word error rate.

ACKNOWLEDGMENT

This paper is based on work supported by the National Science Foundation under grants 01-21285 and 07-03624. The authors would like to thank Ken Chen for making us aware of reference [3].

REFERENCES

[1] K. Livescu, Özgür. Çetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman, S. Dawson-Hagerty, B. Woods, J. Frankel, M. Magimai-Doss, and K. Saenko, "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop," in *Proc. ICASSP*, 2007.

[2] M. Hasegawa-Johnson, K. Livescu, P. Lal, and K. Saenko, "Audiovisual speech recognition with articulator positions as hidden variables," in *International Congress on the Phonetic Sciences*, Saarbrucken, 2007.

[3] G. Hickok and D. Poeppel, "The cortical organization of speech processing," *Nature Reviews*, vol. 8, pp. 393–402, 2007.

[4] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[5] M. Sams, R. Aulanko, H. Hamalainen, O. Lounasmaa, S. Lu, and J. Simola, "Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex," *Neuroscience Letters*, vol. 127, pp. 141–145, 1991.

[6] D. Massaro, *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, NJ: Erlbaum, 1987.

[7] K. Livescu, "Feature-based pronunciation modeling for automatic speech recognition," Ph.D. dissertation, MIT, 2005.

[8] S. Chu and T. S. Huang, "Bimodal speech recognition using coupled hidden Markov models," in *Proc. Interspeech*, Beijing, 2000.

[9] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, pp. 155–180, 1992.

[10] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *IEEE Workshop on ASRU*, Santa Barbara, 1997.

[11] J. L. Elman and J. L. McClelland, "Exploiting lawful variability in the speech wave," in *Invariance and Variability in Speech Processes*, J. Perkell and D. Klatt, Eds. Hillsdale, NJ: Erlbaum, 1986, pp. 360–385.

[12] E. Patterson, S. Gurbuz, Z. Tufecki, and J. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *Proc. ICASSP*, Orlando, 2002.

[13] D. Pallett, "Tools for the analysis of benchmark speech recognition tests," in *Proc. ICASSP*, 1990, pp. 97–100.