BAYESIAN LEARNING FOR MODELS OF HUMAN SPEECH PERCEPTION

Mark Hasegawa-Johnson

ECE Department, University of Illinois; jhasegaw@uiuc.edu

ABSTRACT

Humans speech recognition error rates are 30 times lower than machine error rates. Psychophysical experiments have pinpointed a number of specific human behaviors that may contribute to accurate speech recognition, but previous attempts to incorporate such behaviors into automatic speech recognition have often failed because the resulting models could not be easily trained from data. This paper describes Bayesian learning methods for computational models of human speech perception. Specifically, the linked computational models proposed in this paper seek to imitate the following human behaviors: independence of distinctive feature errors, perceptual magnet effect, the vowel sequence illusion, sensitivity to energy onsets and offsets, and redundant use of asynchronous acoustic correlates. The proposed models differ from many previous computational psychological models in that the desired behavior is learned from data, using a constrained optimization algorithm (the EM algorithm), rather than being coded into the model as a series of fixed rules.

1. INDEPENDENT FEATURE ERRORS

This paper has two goals. First, this paper seeks to introduce a large variety of recent speech psychology results to the statistical signal processing community. Second, this paper proposes computational models of the human behaviors evidenced in these recent results. It is, perhaps, necessary to discuss the reasons why an engineer might be interested in psychology. There is no *a priori* need for automatic speech recognizers to imitate the processes of human speech perception: the experience of the past 30 years indicates that simple but trainable mathematical models consistently achieve lower error rates than psychologically motivated but untrainable models. Despite this progress, however, automatic speech recognition error rates are typically 30 to 300 times worse than human speech recognition error rates [1]. In order to close the gap between human and machine performance, it may be useful to evaluate the performance advantages conferred by specific human speech perceptual behaviors, and to try to imitate the most apparently useful behaviors using trainable machine learning models.

In 1952, Jakobson, Fant and Halle suggested encoding each phoneme as a vector of binary "distinctive features:" voiced vs. unvoiced, lowpass vs. highpass, spectrally compact vs. spectrally diffuse [2]. The idea that a phoneme can be decomposed into independently manipulable dimensions is quite old: classical Greek, Hebrew, Arabic, and Japanese, for example, mark secondary distinctions such as voicing and aspiration by means of diacritics. Jakobson's binary notation was important in part because, within three years after Jakobson's paper, Miller and Nicely were able to prove the psychological reality of a nearly binary distinctive feature notation similar to Jakobson's [3].

Miller and Nicely [3] asked listeners to transcribe noisy recordings of consonant-vowel syllables. Human listeners rarely misunderstand nonsense syllables under quiet listening conditions, but with enough noise, it is possible to get listeners to make mistakes, and the mistakes they make are revealing. First, some distinctive features are more susceptible to noise than others: place of articulation is reliably communicated only at SNR above -6dB, while sonorancy is reliably communicated even at -12dB SNR. Second, errors in the perception of distinctive features are approximately independent, in the following sense: given that the true values of the N distinctive features are $F = [f_1, \ldots, f_N]^T$, the SNR-dependent probability that a listener will perceive the vector $\hat{F} = [\hat{f}_1, \ldots, \hat{f}_N]^T$ is given by

$$p(\hat{F}|F, \text{SNR}) \approx \prod_{i=1}^{N} p(\hat{f}_i|f_i, SNR) \tag{1}$$

Eq. (1) does not specify the dependence of distinctive feature errors on any particular acoustic signal. Several authors have suggested an implementation of Eq. (1) that makes signal-dependence explicit in the following way, where X is the particular acoustic signal used to transmit feature vector F:

$$p(\hat{F}|X) = \prod_{i=1}^{N} p(\hat{f}_i|X) \tag{2}$$

Eq. (2) is motivated by training considerations. Each feature has two possible settings ($f_i = 1$ and $f_i = -1$), thus the feature vector F has 2^N possible settings. A classifier trained to represent $p(\hat{F}|X)$ must distinguish 2^N different labels, while a classifier trained to represent $p(\hat{f}_i|X)$ only distinguishes two labels; the former therefore typically requires 2^{N-1} times as much training data as the latter. Unfortunately, Eq. (2) is incorrect in three ways. First, it is neither a necessary nor sufficient condition for Eq. (1). Second, it is suboptimal as an engineering system: a classifier trained to model $p(\hat{F}|X)$ directly, without factoring as shown in Eq. (2), usually results in fewer errors than a bank of classifiers trained as in Eq. (2). Third, it is not a correct model of human speech perception. Volaitis and

This work was supported by NSF award number 0132900. Statements in this paper reflect the opinions and conclusions of the authors, and are not endorsed by the NSF.

Miller [4], for example, have demonstrated that a voice onset time (VOT) of 40ms is sufficient to turn a synthesized /b/ into /p/, but that /g/ only becomes /k/ when the VOT passes 50ms, i.e. $p(\text{voiced}|X, \text{labial}) \neq p(\text{voiced}|X, \text{velar})$.

2. PERCEPTUAL MAGNET EFFECT

A somewhat better approximation of Eq. (1) may be created by assuming that the perceived feature vector $\hat{F}$ is a deterministic function of the signal X; that is, assume that any given listener will always hear the same sequence of phonemes in response to a given acoustic signal. Specifically, choose any continuous function $G(X) = [g_1(X), \ldots, g_N(X)]$ that specifies the response pattern of listeners by the constraint $\hat{f}_i = \text{sgn}(g_i(X))$. If $G(X)$ is assumed to be a deterministic function, then Eq. (1) is equivalent to

$$p(\hat{F}|F, \text{SNR}) \approx \prod_{i=1}^{N} \int_{\hat{f}_i g_i(X)>0} p(X|f_i, \text{SNR}) dX \tag{3}$$

The function $G(X)$ is, thus far, completely unconstrained, except that $\hat{f}_i = \text{sgn}(g_i(X))$ and Eq. (3) holds. Given these constraints, it is possible to choose $G(X)$ such that the dimensions of $G(X)$ are conditionally independent, i.e.,

$$\int_{\hat{f}_i g_i(X)>0} p(X|f_i, \text{SNR}) dX = \int_0^\infty p(g_i(X)|f_i, \text{SNR}) dg_i \tag{4}$$

where the limits of the right-hand integral are $(0, \infty)$ as shown if $\hat{f}_i = 1$, and $(-\infty, 0)$ if $\hat{f}_i = -1$.

By combining Eq. (3) and (4), a parsimonious speech sound classifier is produced. The classifier consists of two functions: a class-independent multidimensional transform $G(X)$, and a set of class-dependent scalar PDFs $\hat{p}(g_i(X)|f_i)$. The task of a human learner, or of a mathematical model of human speech perception, is to learn functions $G(X)$ and $\hat{p}(g_i(X)|f_i)$ that optimally approximate the unknown PDF $p(X, F)$. Human learners rely primarily on unsupervised learning [5], but methods of optimal learning are only clearly defined for a supervised learner. Specifically, suppose that the learner is given M training tokens of the form (X_m, F_m), $1 \leq m \leq M$, drawn from the unknown PDF $p(X, F)$, where X_m is a waveform and $F_m = [f_{1m}, \ldots, f_{Nm}]^T$ is a vector of distinctive features. The optimal supervised learning algorithm (in the minimum expected K-L divergence sense) is the algorithm that chooses $\hat{p}(g_i(X)|f_i)$ and $g_i(X)$ in order to minimize the empirical cross-entropy,

$$H_{emp}(\hat{p}; p) = -\sum_{m=1}^{M} \sum_{i=1}^{N} \log \hat{p}(g_i(X)|f_i) \tag{5}$$

Cross-entropy measures mismatch between $p(X, F)$ and the modeled PDF $\hat{p}(g_i(X)|f_i)$, but cross-entropy does not measure the classification performance of the resulting model. Recall that the classifier output is $\hat{f}_i = \text{sgn}(g_i(X))$, thus classification errors occur whenever $f_i \neq \text{sgn}(g_i(X))$. Classification error may be improved by subtracting, from the cross-entropy, a regularization term monotonically dependent on the probability of misclassification. For example, if $\hat{p}(g_i|f_i)$ is modeled using a mixture Gaussian model,

$$\hat{p}(g_i|f_i = f) = \sum_{k=1}^{K} c_{ifk} \mathcal{N}(g_i; \mu_{ifk}, \sigma_{ifk}^2) \tag{6}$$

then the simultaneous goals of minimum K-L divergence and minimum classification error may be achieved by maximizing $L(\hat{p}, G)$ for some regularization constant λ:

$$L(\hat{p}, G) = \lambda \sum_{m=1}^{M} \sum_{i=1}^{N} \frac{f_{im}\mu_{ifk}}{\sigma_{ifk}} - H_{emp}(\hat{p}; p) \tag{7}$$

$L(\hat{p}, G)$ may be maximized using the EM algorithm. Given Λ, an initial set of parameters including the parameters of both $G(X)$ and $\hat{p}(g_i(X)|f_i)$, the EM iteration chooses $\hat{\Lambda}$ to maximize

$$Q(\Lambda, \hat{\Lambda}) = \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{k=1}^{K} p(k, g_i(X_m)|f_{im}, \Lambda) \left[\log p(g_i(X_m)|k, f_{im}, \hat{\Lambda}) + \lambda \frac{f_{im}\mu_{ifk}}{\sigma_{ifk}} \right] \tag{8}$$

where k is the mixture index selected to model the ith distinctive feature of the mth training token. Differentiating $Q(\Lambda, \hat{\Lambda})$ with respect to mixture Gaussian parameters results in formulas similar to those in [6]. Differentiating with respect to the parameters of $G(X)$ results in formulas similar to [7].

The acoustic-to-perceptual transform $G(X)$ can be constructed mathematically, but that does not necessarily imply its psychological reality. In order to determine whether or not the space $G(X)$ corresponds to any psychologically real phenomenon, it is necessary once again to examine the psycholinguistic literature.

The ability of listeners to discriminate two nearly identical synthesized speech waveforms (e.g., identical except for a 50Hz difference in the second formant) is highest if the two waveforms straddle a phoneme boundary (e.g., if one waveform is classified as /i/ while the other is classified as /I/). Kuhl and her colleagues [8] have demonstrated that the phoneme boundary does not need to lie between the two waveforms in order to increase their discriminability: two waveforms that are both classified as /i/, but that are both close to the /i/-/I/ boundary, are more discriminable than are two waveforms that are both close to the center of the /i/ region in acoustic space. They explain their results by positing a continuous-valued “perceptual space” computed by the listener as a nonlinear transformation of the acoustic space, $G(X) = [g_1(X), \ldots, g_N(X)]$, such that the magnitude of the Jacobian of the transform is smaller near the center of a phoneme region than it is near the border between phoneme regions [5]. These variations in the value of the Jacobian they term the “perceptual magnet effect.” The proposed perceptual space $G(X)$ is controversial, but continues to serve as an organizing paradigm for new experiments, e.g., [9].

3. REDUNDANT ACOUSTIC CORRELATES

Listeners do not need to hear all of the acoustic evidence for a distinctive feature in order to correctly recognize the feature setting. Phoneticians have catalogued a handful of primary acoustic correlates (characteristic spectrotemporal patterns) that may be used to signal the setting of each distinctive feature. A signal synthesized with any one of these acoustic correlates will be heard to have the target distinctive feature. Consider, for example, the word "backed." This word contains three stop consonants; because of their relative positions in the word, the places of articulation of these three stops are communicated by three very different types of acoustic information. The place of the final /d/ is communicated by a turbulent burst spectrum. The place of the /k/ is communicated by formant transitions during the last 70ms of the vowel. The place of the initial /b/ is communicated by both a turbulent burst and by formant transitions during the first 70ms of the vowel, but experiments with synthetic speech [10] and digitally modified natural speech [11] have shown that either of these cues may be excised without impairing listeners' ability to understand the stop. The closure transition, burst spectrum, and release transition of a stop are thus redundant acoustic correlates; unambiguous presence of any one of these three acoustic patterns is enough to force listeners to hear the desired distinctive feature.

The redundancy principle operates under at least two circumstances. First, one or more acoustic correlates may be missing because of syllable position, as in the example word "backed." Second, one or more acoustic correlates may be inaudible because of noise. When all acoustic correlates are masked by noise, listeners forced to guess the identity of a stop will choose a place of articulation at random. When the noise is lowered sufficiently to unmask either the burst peak or the formant transition, recognition accuracy rapidly approaches 100% [12].

The three sample acoustic correlates discussed above—closure transition, burst spectrum, and release transition—share an important characteristic. All three can only be correctly recognized using a signal representation precisely synchronized with an acoustic-phonetic "landmark:" an instant of sudden signal change, e.g., a consonant closure or consonant release. The mammalian auditory system is uniquely sensitive to sudden onsets and sudden offsets of signal energy [13, 14]. Stevens [15, 16] has proposed a "landmark-based" model of speech perception and recognition, according to which acoustic phonetic landmarks proposed by a pre-processor are then classified by a set of distinctive feature classifiers. Redundancy of asynchronous acoustic observations occurs because landmarks are only classified if they are first detected by the pre-processor, thus if X_1 is a sequence of spectra covering a 140ms period centered at the instant of stop closure, X_2 is a sequence of spectra centered at the stop release, and $\mathcal{X} = [X_1, X_2]$ is their union,

$$p(\mathcal{X}|F) = \begin{cases} p(X_1|F) & \text{if only closure exists} \\ p(X_2|F) & \text{if only release exists} \\ p(X_1|F)p(X_2|F) & \text{if both exist} \end{cases} \tag{9}$$

4. THE VOWEL SEQUENCE ILLUSION

Speech recognition using a landmark-based recognizer requires finding a sequence of landmark times $\mathcal{T} = (t_1, \ldots, t_M)$, and the distinctive feature vectors $\mathcal{F} = (F_1, \ldots, F_M)$ implemented at those landmarks, in order to maximize $p(\mathcal{T}, \mathcal{X}|\mathcal{F})$. Define $\hat{p}(t_m|t_{m-1}, Y)$ to be the probability that a landmark exists at time t_m, given the previous landmark time t_{m-1}, and given an auxiliary observation sequence $\mathcal{Y}$ (e.g., for stop consonant detection, $\mathcal{Y}$ may be a sequence of short-time energies [17]). Then

$$\max_T p(\mathcal{T}, \mathcal{X}|\mathcal{F}) = \max_{t_M} \delta_M(t_M) \tag{10}$$

where

$$\delta_m(t_m) = p(X_m|t_m, F_m) \max_{t_{m-1}} p(t_m|t_{m-1}, \mathcal{Y})\delta_{m-1}(t_{m-1}) \tag{11}$$

Eq. (11) proposes a stochastic landmark-based recognition model (SLM) that may be understood as a new type of stochastic segment model (SSM) [18]. The SLM shares three important properties with the SSM. First, learning algorithms developed for the SSM are also useful for the landmark model, including the segmental K-means algorithm based on Eq. (11), and a Baum-Welch algorithm in which the maxima of Eq. (11) are replaced by sums. Second, as in the SSM, the observations of the landmark model are most naturally defined as spectrogram matrices rather than spectral vectors: specifically, $X_m = [\vec{x}(s_m), \ldots, \vec{x}(e_m)]$ where $\vec{x}(t)$ is the spectral vector computed at time t, s_m is the start time of the mth observation, and e_m is the end time of the mth observation.

Despite these similarities, the landmark model of speech recognition has one advantage that the SSM lacks. Speech recognition using a landmark model may be constrained so that all candidate transcriptions contain exactly the same number of landmarks; as explained in [18], the constraint of equal-length transcriptions is the only condition under which an SSM can make use of explicitly segmental features such as burst spectrum or formant-onset spectrum. Stevens et al. [16] proposed classifying only the landmarks detected by a landmark-detection preprocessor. Not all phoneme landmarks can be accurately detected by a preprocessor, but using a landmark-based recognizer, not all landmarks need to be accurately detected: it is sufficient for the preprocessor to count the number of syllables in the utterance, and to hypothesize that every syllable contains exactly one onset landmark, one nucleus landmark, and one coda landmark. "Relatively reliable" syllable detectors have been available since 1975 [19], and improvements using sub-band periodicity may be possible [20].

Recent evidence from neural physiology and speech perception (e.g., [21]) suggests that human speech perception may depend on a syllable-detection preprocessor, although the evidence for this effect is much weaker than the evidence for redundant acoustic correlates. First, fMRI studies have demonstrated that syllable counting and phoneme recognition occur in different parts of the brain [21]. Second, Warren et al. have demonstrated a "vowel sequence illusion" suggesting that listeners are unable to correctly recognize the phonemes in an utterance unless they are also able to correctly syllabify the utterance [22]. Steady-state vowels, spliced together into a repeating sequence, are easily recognized if each vowel segment is long enough to be a naturally spoken syllable. If the vowel segments are too short to be natural syllables (e.g., 70ms), listeners fail to hear the correct vowels. Instead, listeners hear the signal as a recording of two talkers speaking simultaneously, each talking at a plausible English syllable rate, with phoneme content suggesting that listeners are attributing energy in the high band (above 1500Hz) to one talker, and are attributing energy in the low band (below 1500Hz) to the second talker.

5. CONCLUSION

Four results in the field of speech psychology have been briefly discussed: the independence of distinctive feature errors, the perceptual magnet effect, the redundancy of asynchronous acoustic correlates, and the vowel sequence illusion. Original mathematical models of these behaviors have also been briefly presented; the proposed models are designed to learn these behaviors from standard transcribed speech recognition databases. It remains to be seen whether empirical tests will find the proposed models to be useful as either models of human psychology or as algorithms for automatic speech recognition. Other results relevant to speech psychology, including the many results covered by the heading of "auditory scene analysis" [14], have been omitted for lack of space; future research will seek to incorporate these other psychological results into a parsimonious but complete mathematical learning model of human speech perception.

6. REFERENCES

[1] Richard P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, pp. 1–15, 1997.

[2] R. Jakobson, G. Fant, and M. Halle, "Preliminaries to speech analysis," Tech. Rep. 13, MIT Acoustics Laboratory, 1952.

[3] G. A. Miller and P. E. Nicely, "Analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am*, vol. 27, pp. 338–352, 1955.

[4] Lydia E. Volaitis and Joanne L. Miller, "Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories," *J. Acoust. Soc. Am*, vol. 92, no. 2, pp. 723–735, August 1992.

[5] Frank H. Guenther and Marin N. Gjaja, "The perceptual magnet effect as an emergent property of neural map formation," *J. Acoust. Soc. Am*, vol. 100, pp. 1111–1121, 1996.

[6] Bin H. Juang, Stephen E. Levinson, and Man Mohan Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Trans. on Information Theory*, vol. 32, no. 2, pp. 307–309, 1986.

[7] Mohamed Kamal Omar and Mark Hasegawa-Johnson, "Model enforcement: A unified feature transformation framework for classification and recognition," (in review).

[8] P. K. Kuhl, K. A. Williams, F. Lacerda, K. N. Stevens, and B. Linblom, "Linguistic experience alters phonetic perception in infants by 6 months of age," *Science*, vol. 255, pp. 606–608, 1992.

[9] Anu Sharma and Michael F. Dorman, "Exploration of the perceptual magnet effect using the mismatch negativity auditory evoked potential," *J. Acoust. Soc. Am*, vol. 104, pp. 511–517, 1998.

[10] Pierre C. Delattre, Alvin M. Liberman, and Franklin S. Cooper, "Acoustic loci and transitional cues for consonants," *J. Acoust. Soc. Am*, vol. 27, no. 4, pp. 769–773, July 1955.

[11] Zaki B. Nossair and Stephen A. Zahorian, "Dynamic spectral shape features as acoustic correlates for initial stop consonants," *J. Acoust. Soc. Am*, vol. 89, no. 6, pp. 2978–2991, June 1991.

[12] Abeer A. H. Alwan, *Modeling Speech Perception in Noise: the Stop Consonants as a Case Study*, Ph.D. thesis, MIT, Cambridge, MA, February 1992.

[13] M.I. Miller and M.B. Sachs, "Representation of stop consonants in the discharge patterns of auditory-nerve fibers," *J. Acoust. Soc. Am*, vol. 74, pp. 502–517, 1983.

[14] A. S. Bregman, *Auditory Scene Analysis*, MIT Press, 1990.

[15] K. N. Stevens, "Evidence for the role of acoustic boundaries in the perception of speech sounds," in *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, Victoria A. Fromkin, Ed., pp. 243–255. Academic Press, Orlando, Florida, 1985.

[16] K. N. Stevens, S. Y. Manuel, S. Shattuck-Hufnagel, and S. Liu, "Implementation of a model for lexical access based on features," in *Proc. ICSLP*, Banff, Alberta, 1992, vol. 1, pp. 499–502.

[17] Partha Niyogi, Chris Burges, and Padma Ramesh, "Distinctive feature detection using support vector machines," in *Proc. ICASSP*, Phoenix, AZ, 1999.

[18] Mari Ostendorf, Vassilios V. Digilakis, and Owen A. Kimball, "From HMM's to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 5, pp. 360–378, 1996.

[19] Paul Mermelstein, "Automatic segmentation of speech into syllabic units," *J. Acoust. Soc. Am*, vol. 58, pp. 880–883, October 1975.

[20] Om D. Deshmukh and Carol Y. Espy-Wilson, "A measure of aperiodicity content in a speech signal," *J. Acoust. Soc. Am*, vol. 113, no. 4, pp. 2199, 2003.

[21] Wai Ting Siok, Zhen Jin, P. Fletcher, and Li Hai Tan, "Distinct brain regions associated with syllable and phoneme," *Human Brain Mapping*, vol. 18, no. 3, pp. 201–7, 2003.

[22] Richard M. Warren, Eric W. Healy, and Magdalene H. Chalikia, "The vowel-sequence illusion: Intrasubject stability and intersubject agreement of syllabic forms," *J. Acoust. Soc. Am*, vol. 100, pp. 2452–2461, 1996.