Computer Speech and Language (2001) **15**, 175–194
doi:10.1006/csla.2001.0164
Available online at http://www.idealibrary.com on IDEAL®

A statistical model for robust integration of narrowband cues in speech

Lawrence K. Saul, Mazin G. Rahim and Jont B. Allen

AT&T Labs, 180 Park Ave, Florham Park, NJ 07932, U.S.A.

Abstract

We investigate a statistical model for integrating narrowband cues in speech. The model is inspired by two ideas in human speech perception: (i) Fletcher's hypothesis (1953) that independent detectors, working in narrow frequency bands, account for the robustness of auditory strategies, and (ii) Miller and Nicely's analysis (1955) that perceptual confusions in noisy bandlimited speech are correlated with phonetic features. We apply the model to detecting the phonetic feature [+/−sonorant] that distinguishes vowels, approximants, and nasals (sonorants) from stops, fricatives, and affricates (obstruents). The model is represented by a multilayer probabilistic network whose binary hidden variables indicate sonorant cues from different parts of the frequency spectrum. We derive the Expectation-Maximization algorithm for estimating the model's parameters and evaluate its performance on clean and corrupted speech.

1. Introduction

Two broad lines of research in automatic speech recognition (ASR) are currently motivated by theories of human speech perception. The first, known as *multiband* or *multistream* ASR (Bourlard & Dupont, 1996, 1997; Hermansky, Pavel & Tibrewala, 1996; Tibrewala & Hermansky, 1997; Mirghafori, 1998; Ming & Smith, 1999), is aimed at improving the robustness of machines to noise and corrupting influences in speech; it is inspired by Fletcher's theory of the articulation index (Fletcher, 1953). The second, known as *feature-based* or *acoustic-phonetic* ASR (Deng & Sun, 1994; Espy-Wilson, 1994; Erler & Freeman, 1996; Liu, 1996; Ali, Van der Spiegel, Muller, Haentjens & Berman, 1999; Kirchhoff, 1999; Niyogi, Burges & Ramesh, 1999; King & Taylor, 2000), is aimed at modeling the variabilities introduced by different speakers and linguistic contexts; it is inspired by analyses of sound patterns, models of articulation, and rules of pronunciation (Chomsky & Halle, 1968; Stevens, 1999). In this paper, we hope to demonstrate a synergy between multiband and feature-based ASR, suggesting that certain problems—such as how to integrate narrowband cues in speech—are best tackled by merging these lines of research.

Research in multiband ASR is inspired by a common view of how humans recognize speech (Allen, 1994). In this view, the peripheral auditory system—up through the auditory cortex—appears largely responsible for the resistance of speech to corrupting influences. The cochlea, modeled as a bank of overlapping bandpass filters, resolves incoming signals into components from different parts of the frequency spectrum. Speech is then perceived by

0885–2308/01/020175 + 20 $35.00/0

integrating evidence from different parts of the spectrum, ignoring bands with poor signal-to-noise ratios (SNRs), and capitalizing on bands with informative cues. In this way, humans maintain a robust capacity to recognize speech in poor listening conditions. This view is derived from the pioneering experiments of Fletcher (1953), who showed that listeners can recognize nonsense syllables in corrupted speech, well above chance levels, without the benefit of auxiliary cues. The hope for ASR is that by emulating the multiband processing of the auditory system, we can improve the robustness of automatic methods.

Multiband ASR is based on the hypothesis that different parts of the frequency spectrum should be independently analyzed in the early stages of voice processing. This hypothesis raises the following question: what initial stage of ASR can be attacked by working in "critical" bands of speech; particularly, the narrow frequency bands (roughly between one-third and one-sixth of an octave above about 500 Hz) derived from auditory filters (Moore & Glasberg, 1983)? In particular, if critical bands do not have sufficient bandwidth for robust recognition of whole phonemes, what type of discrete linguistic units (if any) should we be attempting to extract from them?

A natural hypothesis is that multiband processing in ASR should be aimed at the robust detection of phonetic features. Phonetic features categorize phonemes into broad classes based on their articulatory or acoustic properties (Stevens, 1999). Voicing, nasality, and frication are examples of phonetic features. The distinctions implied by these features can be heard in corrupted speech even when listeners cannot identify whole phonemes. This aspect of human performance was quantified in a seminal paper by Miller and Nicely (1955), who measured the mutual information of spoken and perceived phonemes in *noisy bandlimited speech*. Their results show that certain phonetic features are detected early in the speech chain, prior to the general recognition of phonemes, syllables, or words, and that these features are detected more robustly than larger units of speech.

Just as Fletcher's ideas inspired early work in multiband ASR, we believe that Miller and Nicely's results pose many interesting problems in feature detection. Most work in multiband ASR has been devoted to training full-fledged recognizers. The fundamental problem in these recognizers is how to integrate cues from different parts of the frequency spectrum. Feature detection provides a valuable setting to study this fundamental problem without the complications introduced by the massive infrastructure of current recognizers.

In this paper, we propose a statistical model of multiband processing for detecting the phonetic feature [+/−sonorant]. This feature distinguishes vowels, nasals, and approximants (sonorants) from stops, fricatives, and affricates (obstruents). The [+/−sonorant] distinction, though not specifically studied by Miller and Nicely, is extremely robust to corrupting influences in speech (Clark & Yallop, 1995) and is therefore a natural candidate for statistical models of multiband processing. Sonorants are articulated by periodic vibration of the vocal cords with an unobstructed airstream; see Table I for examples. Note that the [+/−sonorant] feature is not to be confused with the attribute of sonority, defined as the loudness of a sound relative to others of the same length, stress, and pitch. The points of greatest sonority in an utterance, measured on a continuous scale, tend to be interpreted as syllable peaks (Clark & Yallop, 1995). This is a different notion than the discrete feature [+/−sonorant] in Table I.

Our method is based on the hypothesis that sonorant cues are detected in narrow frequency bands, and that independent detectors, working in parallel, account for the robustness of auditory strategies. The main contribution of our work is to evaluate this hypothesis in a highly intelligible statistical framework. To this end, we apply modern methods in statistical learning to the lowest level of feature extraction in speech processing. Our front-end consists of an auditory filterbank followed by half-wave rectification and other nonlinearities. Narrowband

TABLE I. The feature [±sonorant] for English consonants. Consonants are further grouped by the feature [±voiced] and the manner of articulation: stop, fricative, affricate, nasal, or approximant

	[+**voiced**]		[−**voiced**]		
[−**sonorant**]	b	(bee)	p	(pea)	*stops*
	d	(day)	t	(tea)	
	g	(gay)	k	(key)	
	z	(zone)	s	(sea)	*fricatives*
	v	(van)	f	(fin)	
	dh	(then)	th	(thin)	
	zh	(azure)	sh	(she)	
	jh	(joke)	ch	(choke)	*affricates*
[+**sonorant**]	m	(mom)			*nasals*
	n	(noon)			
	ng	(sing)			
	l	(lay)			*approximants*
	r	(ray)			
	w	(way)			
	y	(yacht)			

measurements of SNR and periodicity are then fed to a probabilistic graphical model, represented by a multilayer Bayesian network, whose binary hidden variables encode a distributed representation of the speech signal. The structure of the model formalizes a logical hypothesis, inspired by the work of Fletcher (1953), about how to combine information across narrow frequency bands.

Our approach combines a number of previously studied ideas in ASR: multiband processing, detecting phonetic features, hidden variable modeling, and learning from examples. We see these ideas as logically connected in the following way. First, we emulate the multiband processing of the auditory system in an attempt to approach the robustness of human listeners. Second, we study how to detect the feature [+/−sonorant] because larger units of speech (such as phonemes) cannot be recognized as robustly in narrow frequency bands, and because human listeners rarely confuse sonorants with obstruents, even in severely corrupted speech. Third, we adopt hidden variable models because discovering narrowband cues in wideband speech cannot be posed as a problem in fully supervised learning. And finally, we use statistical methods to provide some guarantee of optimality when we fit our hidden variable models to data.

The organization of this paper is as follows. In Section 2, we present our statistical model for detecting the phonetic feature [+/−sonorant]. In particular, we describe its front-end, its graphical representation, and its learning algorithm. In Section 3, we give experimental results for our model in clean and corrupted speech. We also compare these results to the [+/−sonorant] distinction made by standard front-ends and statistical models ASR. Finally, in Section 4, we evaluate the implications of our work for ASR and mention several open problems needing further study.

2. Statistical model

Detecting the phonetic feature [+/−sonorant] can be studied as a problem in statistical pattern recognition (Bishop, 1995). The problem is to design a classifier that takes as input a window of speech, $\mathcal{S}$, and returns as output a conditional probability, $\Pr[+\text{sonorant}|\mathcal{S}]$, between zero and one. The probability measures the certainty that the window mainly overlaps a vowel, approximant, or nasal, as opposed to a stop, fricative, or affricate.

Classifiers of this sort are most commonly trained in the framework of supervised learning. In this framework, windows of speech are labeled as [+/−sonorant] based on phonetic alignments of a large corpus, and a learning algorithm is used to optimize an objective function, such as the average error rate or log likelihood. Classifiers trained in this way can accurately distinguish between sonorants and obstruents in clean speech. Without further constraints, however, they do not generalize well to noisy or filtered speech whose characteristics do not precisely match the training data. This lack of robustness—typical of baseline systems in ASR—motivates our search for richer statistical models that incorporate ideas in multiband processing.

A multiband model of [+/−sonorant] detection must satisfy additional criteria. The input to such a model is speech passed through a filterbank. Let $\mathcal{S}_i$ denote a window of bandlimited speech from the ith frequency band. A multiband model must predict whether $\mathcal{S}_i$, by itself, contains a cue for [+sonorant] speech. In addition, the model must compute a wideband probability, $\Pr[+\text{sonorant}|\mathcal{S}_1, \mathcal{S}_2, \ldots]$, that integrates narrowband cues from different parts of the frequency spectrum.

This type of classification poses a challenging problem in machine learning because there are no labeled examples of bandlimited speech. In particular, while phonetic alignments provide a [+/−sonorant] segmentation of wideband speech, they do not provide a frequency profile of [+sonorant] segments, thereby indicating which parts of the spectrum contain [+sonorant] cues and which do not. We overcome this problem of unlabeled (or "missing") data by treating the narrowband cues as hidden variables in a larger statistical model. This enables us to derive an Expectation-Maximization (EM) algorithm (Dempster, Laird & Rubin, 1977) for estimating the parameters of the model from phonetically segmented speech.

The description of this model is divided into four parts. First, we describe the front-end and the acoustic measurements extracted from the speech waveform. Second, we describe the model architecture in general terms, explaining at a high level how it manages to balance competing criteria for robustness. Third, we show how to represent the model by a Bayesian network (Pearl, 1988). This enables us to compute the statistics of hidden variables using message-passing algorithms for probabilistic inference. Finally, we present the EM algorithm for parameter estimation.

2.1. *Multiband processing*

In sonorant speech, we expect energy to be concentrated at harmonics (equally spaced multiples of the speaker's pitch). When such speech is degraded by noise or filtering, we expect periodicity cues to survive in parts of the spectrum having a high SNR. Our front-end therefore focuses on two acoustic correlates of sonorant speech: (i) the periodicity established by the speaker's pitch (Hess, 1983; Holmes, 1998), and (ii) increases in the SNR (or in the case of clean speech, the signal-to-background ratio). These acoustic correlates can be observed in narrow frequency bands. Our general approach to sonorant detection is inspired by the phenomenon of *residue pitch* (Moore, 1997), the well-known effect that listeners can perceive pitch from different parts of the frequency spectrum. Many of the specific operations in our front-end were adapted from biologically motivated models of pitch processing (Slaney & Lyon, 1990; Smith, 1996, 1997).

Our front-end begins by transforming the speech waveform into an array of narrowband envelopes. These envelopes are computed by half-wave rectifying, squaring, lowpass filter-

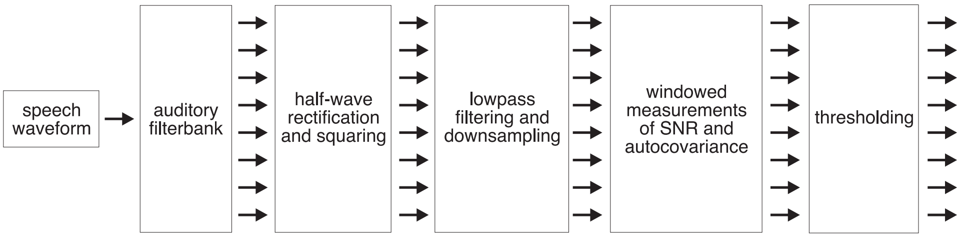
speech waveform
auditory filterbank
half-wave rectification and squaring
lowpass filtering and downsampling
windowed measurements of SNR and autocovariance
thresholding

Figure 1. Stages of front-end processing.

ing, and downsampling the outputs of auditory filters with center frequencies between 225 and 3625 Hz. Parallel measurements are made on sliding windows of these envelopes. Six measurements are made per frame per critical band: the first two are running estimates of the SNR, while the remaining four are autocovariance statistics. The measurements are normalized by threshold values derived from identically processed bands of white noise. The overall scheme is illustrated in Figure 1; a more detailed description is given in Appendix A. There are 24 channels, resulting in a total of 144 measurements per 16 ms window of speech. Just over 60 frames are processed per second, with contiguous nonoverlapping windows. These frames are fed, one at a time (or in small groups), to the statistical model described in the next section.

2.2. Balancing errors: false positives vs. false negatives

A robust [+/−sonorant] detector must balance competing demands of selectivity and invariance. On one hand, to avoid false positives, it must be selective about what constitutes evidence for [+sonorant], focusing on appropriate measures of periodicity and SNR, and discounting incidental correlations that also happen to be generated by noise. On the other hand, to avoid false negatives, it must be sensitive to a variety of cues, monitoring the entire frequency spectrum for significant acoustic events, and not relying too much on any one particular cue (e.g. low-frequency energy) that can be destroyed by noise or filtering.

Our model has two stages that explicitly address these concerns. To convey the basic intuition behind these stages, we will first describe them in terms of simple binary events and logical decision rules. In the next section, however, we will model the binary events as random variables and show how to propagate uncertainty in their outcomes throughout the entire decision-making process.

Roughly speaking, then, in the first stage of the model, independent assessments of [+/−sonorant] are made on sliding windows of bandlimited speech generated by the filterbank. A window of bandlimited speech is labeled as [+sonorant] if its measurements of SNR and autocovariance meet a number of criteria. Effectively, these criteria are treated as inputs to a logical AND gate. Only by meeting all the criteria is the AND condition satisfied and the window of bandlimited speech labeled as [+sonorant]. The requirement that several criteria must be simultaneously satisfied is designed to minimize errors due to false positives.

In the second stage of the model, an overall assessment of the feature [+/−sonorant] is made for the frame of wideband speech. A simple prescription is used to combine information across frequency bands. In particular, the wideband speech is labeled as [+sonorant] if one or more frequency bands are labeled as [+sonorant]. Note that this rule amounts to feeding the outputs of the previous AND gates to a logical OR gate. Only when no individual frequency band is labeled as [+sonorant] is the OR condition violated and the wide-

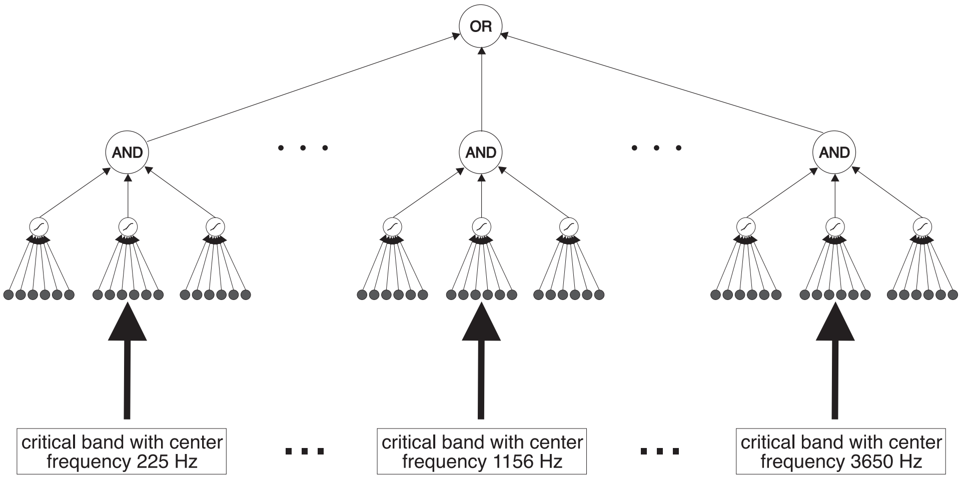
OR
AND
AND
AND
critical band with center frequency 225 Hz
critical band with center frequency 1156 Hz
critical band with center frequency 3650 Hz

Figure 2. Multilayer Bayesian network for detecting the feature [+/−sonorant]. The inputs are windowed measurements from narrow frequency bands of speech.

band speech labeled as [–sonorant]. Thus, as long as periodicity cues are detected in some part of the spectrum with a high SNR, the frame of speech is labeled as [+sonorant]. This way of combining information is motivated by Fletcher's hypothesis (Fletcher, 1953; Allen, 1994) that independent detectors, working in parallel, account for the robustness of auditory strategies. A similar model for combining narrowband cues in ASR was also investigated by Ming and Smith (1999).

The logical operations in this model represent consensus strategies for balancing the demands of selectivity and invariance. AND gates, triggered only by a consensus of positive inputs, work to minimize false positives. OR gates, silenced only by a consensus of negative inputs, work to minimize false negatives. Together, these operations help a [+/–sonorant] detector trained on clean speech to generalize well to noisy bandlimited speech. In particular, while the AND operations prevent false positives from noise, the OR operation makes it unlikely that evidence of [+sonorant] will be completely missed, even when most of the frequency spectrum has been corrupted.

2.3. Probabilistic graphical model

Figure 2 depicts the statistical model described in the previous section as a multilayer Bayesian network (Pearl, 1988). The nodes in this network represent random variables, while the links represent statements of conditional dependence. The bottom–up flow of the network describes the sequence of operations used to determine if a frame of speech is [+/–sonorant]. The model is also used to propagate uncertainties in the outcomes of the AND and OR operations. Note that the direction of the arrows in the network is from bottom-to-top: thus, the network parameterizes a discriminative model for classifying [+/–sonorant] frames of speech, as opposed to a generative model for evaluating the likelihood of acoustic measurements.

The nodes in the bottom layer represent the six measurements of SNR and autocovariance (per frame) in each critical band. The variables represented by these nodes are always instantiated; that is, determined by measurements of the waveform, whereas the other nodes in the network represent binary (0/1) random variables with genuine uncertainty. The nodes in the bottom layer are shaded to emphasize this distinction.

The nodes in the second layer represent binary variables indicating whether the measurements of SNR and autocovariance meet certain criteria for detecting [+sonorant]. We use $\boldsymbol{M}_i$ to denote the vector of acoustic measurements in the ith critical band, and X_{ij} to denote whether or not the jth criterion in this band is satisfied. The conditional probability that this criterion is satisfied is given by:

$$\Pr[X_{ij} = 1|\boldsymbol{M}_i] = \sigma(\boldsymbol{\theta}_{ij} \cdot \boldsymbol{M}_i), \tag{1}$$

where $\sigma(z) = [1 + e^{-z}]^{-1}$ is the logistic function. The nodes in the second layer are marked by sigmoids to indicate the logistic function in Equation (1). The weights $\boldsymbol{\theta}_{ij}$ in this equation are parameters in a logistic regression; they are estimated by automatic methods, as described in Section 2.4. Though not explicitly indicated, a default input of unity can be appended to the measurement vector $\boldsymbol{M}_i$ in order to accommodate a bias term in the logistic regression.

The nodes in the third layer represent the assessments of [+/−sonorant] in each critical band. We use Y_i to denote the binary random variable for the ith critical band. As described earlier, a positive assessment is made only if all the criteria X_{ij} are satisfied. Thus, the conditional probability distribution for Y_i is given by:

$$\Pr[Y_i = 1|\boldsymbol{M}_i] = \prod_j \Pr[X_{ij} = 1|\boldsymbol{M}_i]. \tag{2}$$

The nodes in this layer are labeled by AND to indicate the conjunction relating X_{ij} to Y_i.

The node in the top layer represents the overall assessment of [+/−sonorant] for the frame of wideband speech. We use Z to denote this binary random variable. As described earlier, the frame is labeled as [+sonorant] if one or more critical bands are labeled as [+sonorant]. Thus, the conditional probability distribution for Z is given by:

$$\Pr[Z = 1|\boldsymbol{M}] = 1 - \prod_i (1 - \Pr[Y_i = 1|\boldsymbol{M}_i]), \tag{3}$$

where we have used $\boldsymbol{M} = \{\boldsymbol{M}_1, \boldsymbol{M}_2, \ldots\}$ as shorthand to denote the entire set of measurements. The top node in the network is labeled by OR to indicate the disjunction relating Y_i to Z.

For [+sonorant] frames of speech, we can recast Equation (3) in a more familiar form. In this case, the probability of error is given by:

$$1 - \Pr[Z = 1|\boldsymbol{M}] = \prod_i (1 - \Pr[Y_i = 1|\boldsymbol{M}_i]). \tag{4}$$

This is Fletcher's product-of-errors rule (Fletcher, 1953; Allen, 1994). In this context, the rule states that the overall error rate for mistaking obstruents as sonorants is equal to the product of error rates from narrowband detectors.

So far we have described how this model computes $\Pr[Z|\boldsymbol{M}]$—namely, the probability that a frame of speech is [+/−sonorant] based on narrowband measurements of SNR and periodicity. This inference involves a bottom–up propagation of information through the network in Figure 2. Other inferences can also be made, involving a combination of bottom–up and top–down reasoning. Posterior probabilities, such as $\Pr[X_{ij}|Y_i, \boldsymbol{M}_i]$ and $\Pr[Y_i|Z, \boldsymbol{M}]$, are of particular interest for the problem of learning from examples. Certain of these posterior probabilities follow trivially from the AND and OR operations. For example, based on the AND operation, we can make the inference $\Pr[X_{ij} = 1|Y_i = 1, \boldsymbol{M}_i] = 1$, or if $Y_i = 1$, then $X_{ij} = 1$ for all j. Likewise, based on the OR operation, we can make the inference $\Pr[Y_i = 1|Z = 0, \boldsymbol{M}] = 0$, or if $Z = 0$, then $Y_i = 0$ for all i. Other posterior probabilities can be computed from Bayes rule. To simplify the resulting expressions, we use

$p_{ij} = \Pr[X_{ij} = 1|\boldsymbol{M}_i]$ to denote the conditional probabilities computed by Equation (1) in the bottom layer of the network. In [–sonorant] frequency bands, we can make the inference:

$$\Pr[X_{ij} = 1|Y_i = 0, \boldsymbol{M}_i] = p_{ij}\left[\frac{1 - \prod_{k \neq j} p_{ik}}{1 - \prod_l p_{il}}\right]. \tag{5}$$

The term in square brackets on the right-hand side of this equation is always less than one. Thus Equation (5) states that when one or more criteria in the bottom layer is not satisfied, we should decrease our belief that any particular criterion is satisfied. Likewise, in [+sonorant] frames of speech, we can make the inference:

$$\Pr[Y_i = 1|Z = 1, \boldsymbol{M}] = \frac{\prod_k p_{ik}}{1 - \prod_m \left(1 - \prod_l p_{ml}\right)}. \tag{6}$$

The denominator in this equation is always less than one. Thus Equation (6) states that when [+sonorant] speech has been detected in one or more critical bands, we should increase our belief that it was detected in any particular critical band. The advantage of the probabilistic graphical model is that it formalizes these intuitions in a quantitatively precise way.

Two simple extensions to the above model are useful for producing smooth [+/−sonorant] estimates as a function of time. The first is to include first- and second-order time derivatives of the SNR and autocovariance statistics in the measurement vector. The second is to feed measurements from consecutive frames (as opposed to the same frame) to the logistic regressions under each AND gate. These extensions do not complicate the inference procedures in any way. Both were used in our experiments to improve the overall performance of the model.

Again, it is worth noting that certain types of inferences cannot be made from the network in Figure 2. We emphasize that the bottom nodes in the network are always assumed to be instantiated, and that the arrows in the network point from bottom to top. Thus, the network defines a discriminative model, from which to compute $\Pr[Z|\boldsymbol{M}]$, but not a generative model, from which to sample $\Pr[\boldsymbol{M}|Z]$. Thus, the network cannot be used to evaluate the likelihood of acoustic measurements or to fill in missing data (Cooke, Green, Josifovski & Vizinho, 2001). The purely bottom–up aspect of the model's computation has both advantages and disadvantages; these are discussed more fully in Section 4.2.

2.4. Learning algorithm

The quantitative predictions of the model in Figure 2 are determined by the values of its parameters. These parameters—the weights $\boldsymbol{\theta}_{ij}$ in the bottom layer of the network—must be estimated from training data. The training data for the network consists of frames of wideband speech, labeled as [+/−sonorant] based on phonetic alignments.

The parameters of the model are tuned so that its [+/−sonorant] predictions match (with high probability) the labels indicated by the phonetic alignment. Specifically, to each frame indexed by the superscript t, we associate a set of acoustic measurements, $\boldsymbol{M}^t$, and a target label, $z^t \in \{0, 1\}$, indicating whether or not the frame is [+sonorant]. The model parameters are found by attempting to minimize the cross entropy error function (Bishop, 1995):

$$\mathcal{E} = -\sum_t \{z^t \log \Pr[Z^t = 1|\boldsymbol{M}^t] + (1 - z^t)\log(1 - \Pr[Z^t = 1|\boldsymbol{M}^t])\}, \tag{7}$$

whose two terms sum over [+sonorant] and [−sonorant] frames of speech. Due to the chaining of AND and OR operations in Equations (2) and (3), this error function depends in a highly nonlinear way on the weights $\boldsymbol{\theta}_{ij}$ in the bottom layer of the network. The internal structure of the model, however, can be exploited to derive a simple learning algorithm. This is done via the EM algorithm (Dempster *et al.*, 1977), a general iterative procedure for parameter estimation in hidden variable models.

The EM algorithm consists of two alternating steps, an E-step and an M-step. The E-step in this model is to compute the posterior probabilities $\Pr[X_{ij}|Z, \boldsymbol{M}]$, conditioned on the labels provided by the phonetic transcription. The M-step is to update the parameters in each logistic regression, using these posterior probabilities as target values. The value of this two-step approach is to decouple the original problem of minimizing the log loss into several independent weighted logistic regressions. By a general convergence theorem (Dempster *et al.*, 1977), each iteration of these steps is guaranteed to decrease the overall log loss. Details of the algorithm are given in Appendix B.

The EM algorithm for this model solves a version of the "multiple-instance" learning problem (Maron & Lozano-Perez, 1998); namely, it trains a classifier from collections of examples that are ambiguously labeled. In multiple-instance learning, a collection of examples is labeled as negative if all the examples in it are negative, or as positive if one or more examples in it are positive. The problem of detecting [+sonorant] cues in narrowband speech fits neatly into this framework. In our case, the phonetic alignment labels the wideband speech as obstruent (negative) or sonorant (positive). A negative label indicates that no [+sonorant] cues exist in individual critical bands: $\Pr[Y_i = 0|Z = 0] = 1$ for all i. On the other hand, a positive label indicates that [+sonorant] cues exist in one or more bands: $\Pr[\sum_i Y_i \geq 1|Z = 1] = 1$. The EM algorithm solves the problem of multiple-instance learning by inferring individual labels from the posterior distribution, $\Pr[Y_i|Z, \boldsymbol{M}]$.

3. Experiments

We conducted experiments on the TIMIT speech corpus (Garofolo, 1988), whose phonetic transcriptions and speech waveforms have been manually aligned. Phonetically derived segmentations of [+/−sonorant] speech were used to train the multiband model described in the previous section. Frames of speech were labeled as [+sonorant] if they were predominantly aligned with sonorants, and as [−sonorant] if they were predominantly aligned with obstruents or silence. Except for three special cases, vowels, nasals, and approximants were treated as sonorants, and stops, fricatives, and affricates as obstruents. The three special cases were: flapped /d/ ("ladder"), which was treated as [+sonorant], and voiceless /h/ ("hay") and devoiced schwa ("suspect"), which were treated as [−sonorant]. These exceptions can be viewed as borderline cases. The flapped /d/ is a rapidly articulated voiced stop that does not interrupt the periodicity of its neighboring sonorants. Likewise, the voiceless /h/ and devoiced schwa—though articulated with an unobstructed airstream—do not contain periodic acoustic energy. The [+/−sonorant] labels derived in this way should not be construed as absolute truth, but rather as imperfect (yet generally consistent) targets for the training and evaluation of frame-based statistical models.

We also investigated the [+/−sonorant] distinction made by traditional acoustic models in ASR (Rabiner & Juang, 1993). This was done to assess how robustly current models capture the same bit of phonetic information. Two Gaussian mixture models (GMMs)—one for [+sonorant] speech, one for [−sonorant] speech—were trained from windowed (16 ms) measurements of mel-frequency cepstra and log-energy. First- and second-order time derivatives were included in the cepstral feature vectors, and feature variability was reduced by utterance-based energy normalization and cepstral mean subtraction. The posterior probabilities from

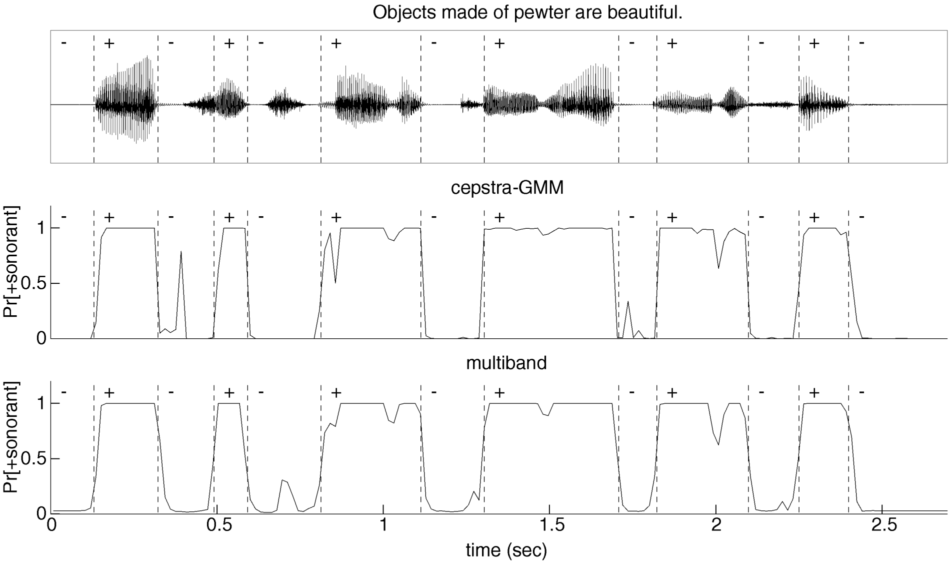

Figure 3. Top: manual segmentation of clean speech into [+/−sonorant] segments. Middle and bottom: probabilities Pr[+sonorant] from cepstra-GMM and multiband classifiers.

these GMMs were used to classify frames of speech as [+/−sonorant]. The parameters of the GMMs were estimated by an EM algorithm (Bishop, 1995). Each GMM had 32 mixture components with diagonal covariance matrices. The number of mixture components was chosen to optimize the error rate in a matched testing condition.

The multiband and cepstra-GMM classifiers were trained exclusively on clean wideband speech sampled at 8 kHz. Overall, the multiband and cepstra-GMM classifiers had 1368 and 5056 parameters, respectively, which had to be estimated from training data. Training data consisted of 380 sentences, or 112 586 frames of speech, nearly evenly divided between [+sonorant] and [−sonorant] frames. Testing data consisted of 110 sentences from different speakers. All speakers were taken from the first dialect region of the TIMIT corpus; this was done simply to limit the size of the experiments in a controlled way. After training, we evaluated the robustness of the multiband and cepstra-GMM classifiers by measuring the frame error rates in a wide variety of listening conditions. SNRs for additive background noise were computed from the total energy ratios of speech to noise over non-silent (endpointed) regions of the speech waveforms.

3.1. Example

Our results are best introduced by considering an illustrative example. Figure 3 shows the waveform for the test utterance “OBJECTS MADE OF PEWTER ARE BEAUTIFUL” spoken in quiet. Dashed vertical lines indicate the boundaries between [+/−sonorant] segments, as determined by the manual alignment. The bottom plots in this figure show the probabilities Pr[+sonorant] for consecutive frames, as computed by the cepstra-GMM and multiband classifiers. [See Equation (3) for the latter inference.] Note that in quiet surroundings, both models provide nearly perfect segmentations of this utterance. In contrast, Figure 4 shows the same utterance contaminated by 0 dB bandlimited noise from 0 to 1 kHz. Despite the

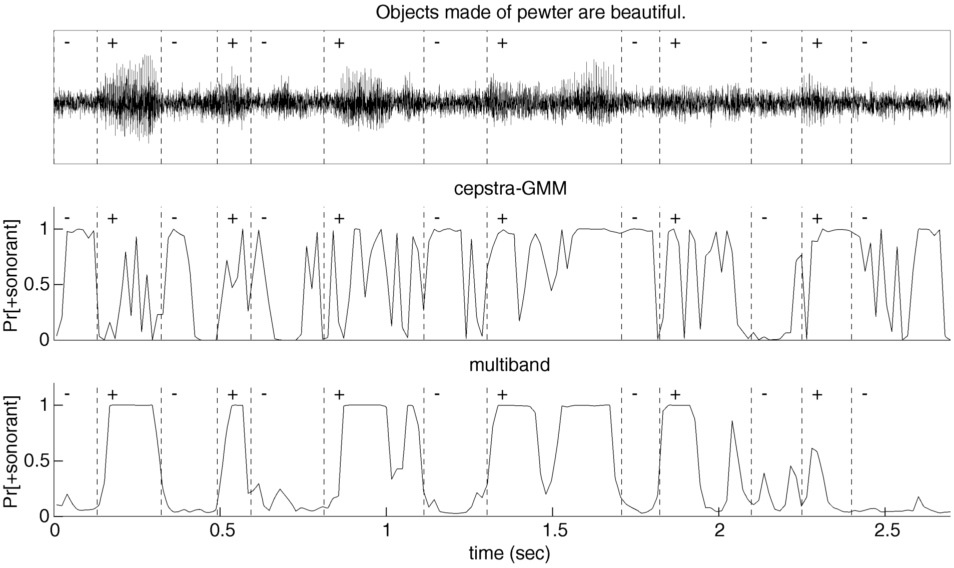

Figure 4. Top: manual segmentation of noisy speech into [+/−sonorant] segments. The speech was contaminated by 0 dB bandlimited white noise from 0 to 1 kHz. Middle and bottom: probabilities Pr[+sonorant] from cepstra-GMM and multiband classifiers.

ragged appearance of the waveform, the speech in this example is quite intelligible. The cepstra-GMM classifier, however, completely breaks down because the cepstral feature vectors conflate different parts of the frequency spectrum. On the other hand, the multiband classifier degrades gracefully, focusing on those critical bands with high SNR.

Figure 5 illustrates how the multiband model monitors sonorant activity in different critical bands for these waveforms. Recall that the hidden variables Y_i in the multiband model encode a frequency profile of sonorant cues in the speech signal. The plots in Figure 5 show the probabilities $\Pr[Y_i = 1|\boldsymbol{M}_i]$ for the clean and noisy waveforms, as computed from Equation (2). The center frequencies of the critical bands are shown to the left of the y-axes. Note that for each frame of speech, the multiband model produces a complete profile of sonorant activity by frequency, as opposed to a single wideband measure. Appropriately, for the noisy waveform, almost all the sonorant activity is detected above 1000 Hz. The plots unambiguously illustrate that the hidden variables in our model have learned to represent frequency-dependent phonetic cues.

3.2. Results

These patterns of generalization were also observed in more formal experiments. We evaluated the multiband and GMM classifiers in 10 testing conditions: (i) wideband speech in quiet, (ii) telephone distortion from the nTIMIT database (Jankowski, Kalyanswamy, Basson & Spitz, 1990), (iii) 0 dB white noise, (iv)–(vii) 0 dB bandlimited noise from 0 to 1 kHz, 1 to 2 kHz, 2 to 3 kHz, and 3 to 4 kHz, and (viii)–(x) speech bandlimited to 0–1 kHz, 1–2 kHz, and 2–3 kHz.

Performance was measured by frame error rates for detecting [+/−sonorant] speech. These error rates measure the percentage of incorrectly labeled frames, including both sonorant–

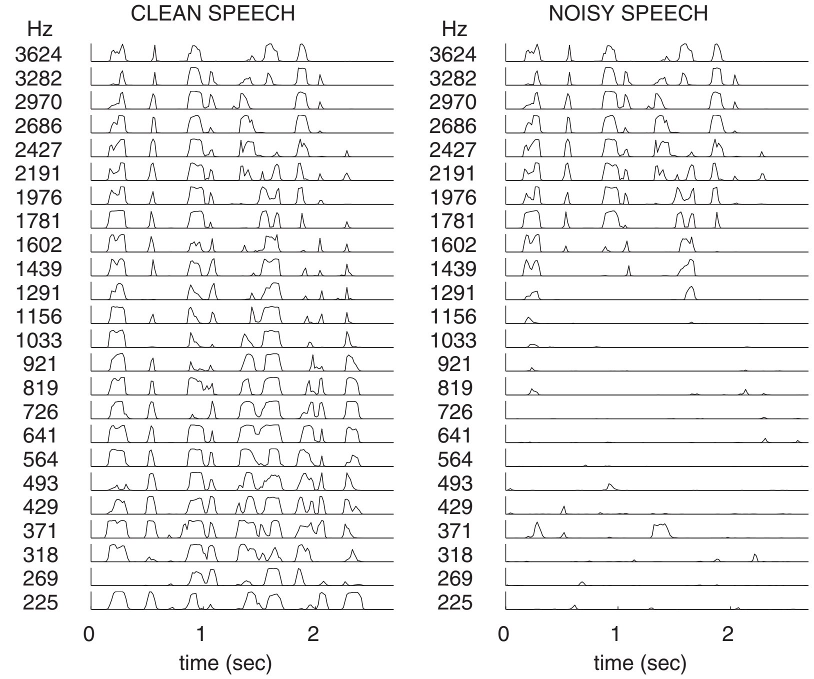

Figure 5. Sonorant activity in critical bands with different center frequencies for the waveforms in Figures 3 and 4. The plots show the probabilities $\Pr[Y_i = 1|\boldsymbol{M}_i]$ of hidden variables in the multiband classifier, as computed from Equation (2).

obstruent and obstruent–sonorant confusions. The top panel of Figure 6 compares these error rates for the multiband and cepstra-GMM classifiers. Roughly speaking, given the number of test frames—over 20 000—we believe that differences in error rates greater than $1/\sqrt{20\,000} \approx 0.7\%$ can be regarded as statistically significant. Applying more stringent criteria, based on the number of phonemes or the number of [+/−sonorant] segments in the test set, leads to more conservative estimates for statistically significant differences in error rate. These estimates are 1.5%, based on the number of phonemes, and 2.3%, based on the number of [+/−sonorant] segments. Using any of these criteria, however, we can draw a number of conclusions.

As expected, the multiband model is significantly more robust to bandlimited noise. The superiority is particularly marked for low frequency noise below 1 kHz. Note that the bandlimited noise in these experiments had no special relation to the auditory filters in our front-end.

In the other test environments, the differences in overall error rates do not appear as significant. Note, however, that sonorant cues are generally destroyed by noise and filtering, and thus one expects higher error rates—arising from false negatives—in these types of test conditions. It is therefore instructive to compare false positive rates, which measure only the errors in [−sonorant] frames that do not contain vowels, nasals, or approximants. The bottom panel of Figure 6 compares false positive rates. The multiband model produces significantly fewer false positives than the cepstra-GMMs; in many cases, the number is lower by an order of magnitude. A revealing trend is that severe noise and filtering, which destroy periodicity cues in the actual speech signal, actually lead to higher numbers of false positives in the

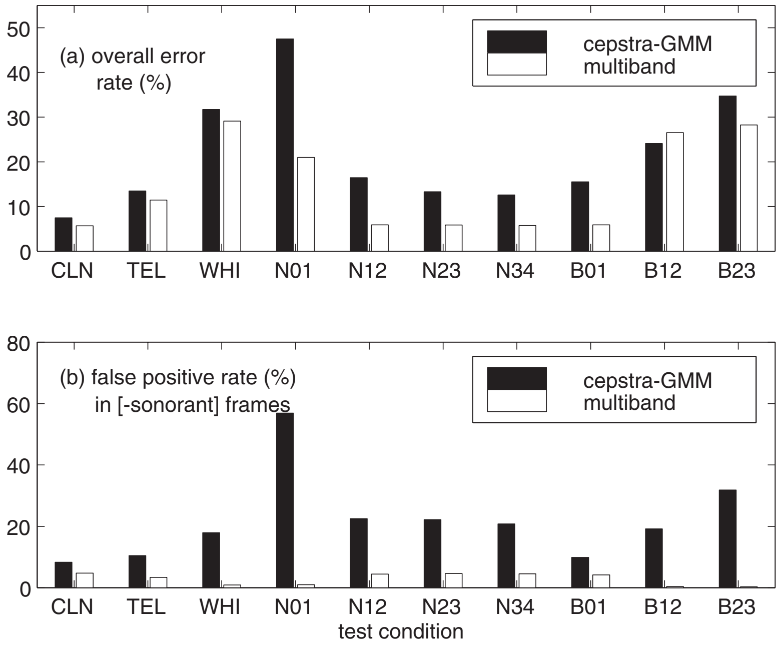

Figure 6. Frame error rates for [+/−sonorant] detection in 10 different environments: clean (CLN), telephone channel (TEL), white noise (WHI), bandlimited noise (N f_1 f_2), and bandlimited speech (B f_1 f_2), where f_1 and f_2 are measured in kHz.

cepstra-GMMs. Thus it appears that the cepstra-GMMs learn to cue on acoustic events that are incidentally correlated with sonorants in clean speech (such as energy bursts at low frequency), but completely uncorrelated with sonorants in general. These results shed light on the fragility of current recognizers.

We also investigated the [+/−sonorant] detection as a function of SNR. Four types of broadband noise were added to the 110 test utterances at SNRs ranging from 30 dB to 0 dB. The noises—babble, factory floor, white, and pink—were taken from the NOISEX database (Varga, Steeneken, Tomlinson & Jones, 1992). Figure 7 shows the frame error rates of the cepstra-GMM and multiband classifiers vs. SNR. The results indicate that the multiband model is more robust over a wide range of SNRs.

The above results demonstrate the viability of the multiband approach. The numbers in Figures 6 and 7, however, are less important than the overall picture of multiband processing that emerges from Figure 5. As we discuss below, the multiband model in this paper has many shortcomings, even for the narrowly defined task of [+/−sonorant] detection. To the best of our knowledge, however, it represents the first statistical model of its kind: a multilayer Bayesian network—purposefully structured to encode Fletcher's product-of-errors rule—whose hidden variables monitor narrowband phonetic cues, and whose parameters are jointly estimated to optimize a wideband measure of performance.

4. Discussion

The multiband processing in our model provides a certain degree of robustness to many types of noise and distortion. To avoid "misses", or false negatives, the model exploits the

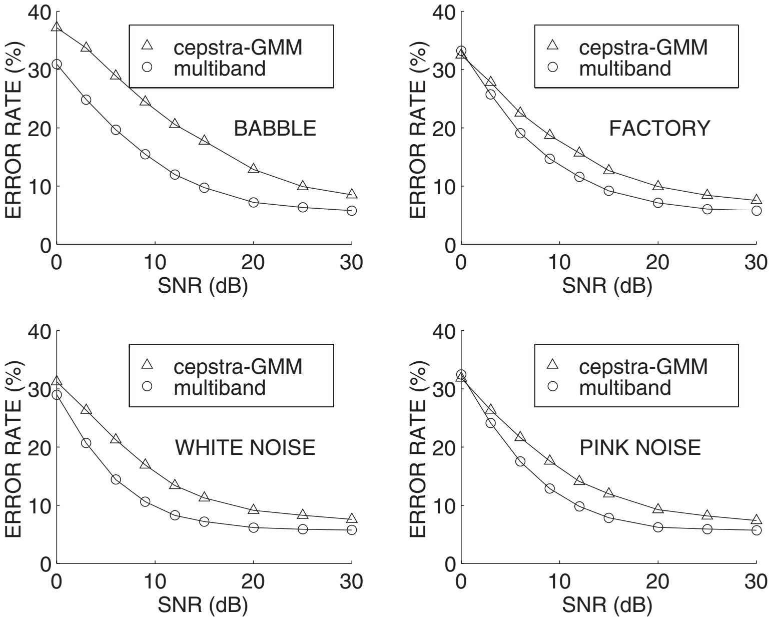

Figure 7. Frame error rates for [+/−sonorant] detection as a function of SNR.

idea that evidence for [+sonorant] is distributed across the frequency spectrum. Likewise, to avoid false positives, the model looks (within each critical band) for multiply consistent measurements of periodicity and SNR. These measurements are derived from auditory filters and purposeful nonlinearities, as opposed to smoothed power spectra and cepstra, which tend to lose their predictive value in corrupted speech. Though the model is trained only on clean wideband speech, its performance degrades gracefully in noisy bandlimited speech.

To what does the model owe its robustness—the nature of its signal processing, or the structure of its decision making? In related experiments, we attempted to train an unstructured classifier, with the same number of free parameters, that took as input the same narrowband measurements of the waveform but did not encode the AND-OR integration of narrowband cues. We found the training of such a classifier, whose input vector is formed by concatenating all the inputs to the bottom layer of the network in Figure 2, to be considerably more difficult due to the high dimensionality of its input—nearly two orders of magnitude greater than that of the individual narrowband detectors. Thus, it should be emphasized that the structure of the AND-OR network not only serves to incorporate prior knowledge, but also to decompose the learning problem into simpler, independent sub-processes, as discussed in Section 2.4 and Appendix B. Based on our experience, it seems unlikely that a classifier whose learning procedure does not exploit such a decomposition could realize the full potential of multiband signal processing. (It is also worth pointing out that such a classifier could only function as a "black box" for [+/−sonorant] detection, whereas the hidden variables in the multiband model can in principle support other useful inferences for ASR.) Finally, we note that the unstructured classifier did not learn to recognize [+/−sonorant] cues from different parts of the frequency spectrum, so that its performance degraded catastrophically in highpass filtered speech.

4.1. Implications for ASR

Today's speech recognizers suffer from a tremendous sensitivity to the acoustic environment; in particular, moderate amounts of noise or frequency response distortion can lead to catastrophic failures. Current systems typically couple a back-end of continuous density hidden Markov models (HMMs) to a front-end that computes mel-frequency (or similarly smoothed) cepstra (Rabiner & Juang, 1993). Viewed as a long-term strategy for matching human performance (Lippmann, 1997), this program for ASR has inherent limitations. Traditional back-ends do not provide much support for emulating the multiband processing of the auditory system. The smoothing of the spectra suppresses the periodic signature of the speaker's pitch, blurring the elementary phonetic distinctions that can be made on this basis. Finally, the use of cepstra conflates different parts of the frequency spectrum, focusing on incidental energy ratios instead of robust phonetic cues. The results of the cepstra-GMMs in the previous section reflect these significant limitations.

Our results support the idea that emulating the multiband processing of the auditory system can improve the robustness of automatic methods. As we have seen, this approach to robustness gives rise to an important tradeoff involving bandwidth. How narrow should we make the filters in the front-end? On one hand, the filters should be sufficiently narrow to reject noise in large parts of the spectrum. On the other hand, the filters should be sufficiently wide to preserve useful phonetic cues. We believe one way to balance these competing goals is to work in critical bands and look for phonetic features.

4.2. Relation to previous work

The most ambitious work in multiband ASR has led to full-fledged recognizers (Bourlard & Dupont, 1996, 1997; Hermansky *et al.*, 1996; Tibrewala & Hermansky, 1997; Mirghafori, 1998; Ming & Smith, 1999). In this paper, we have attempted to detect the feature [+/−sonorant] from critical band measurements of SNR and periodicity. In contrast, existing multiband recognizers tend to work in much wider frequency bands, searching for larger units of speech based on subband cepstra. While these recognizers have demonstrated improvements in robustness, we believe that the multiband approach can be pushed further than this, and that long-term progress requires greater understanding of narrowband phonetic cues. The difficulty of this approach, which we acknowledge, is that progress cannot be so easily measured in terms of word error rates.

Another compelling approach to robust ASR treats noisy frequency bands as missing data (Cooke *et al.*, 2001), making use of probabilistic methods—such as marginalization or imputation—to handle acoustic evidence from unreliable parts of the spectrum. These methods can be viewed as forms of top–down reasoning, since they involve a generative model of acoustic measurements that are correlated across frequency and time. The power of the missing data approach is that it can exploit these correlations to make inferences about corrupted parts of the frequency spectrum. Nevertheless, it remains of interest to study whether purely bottom–up computations, which are considerably faster, can lead to robust phonetic discriminations. Our model, in which independent detectors work in narrow frequency bands, has the advantage that its computations are relatively cheap; thus, at least for [+/−sonorant] it may provide a reasonable degree of robustness at less computational cost. The disadvantage of this approach is that it does not exploit the potential of top–down reasoning.

Feature detection continues to be an active area of research in ASR (Deng & Sun, 1994; Espy-Wilson, 1994; Erler & Freeman, 1996; Liu, 1996; Ali *et al.*, 1999; Kirchhoff, 1999; Niyogi *et al.*, 1999; King & Taylor, 2000). Our approach to [+/−sonorant] detection is based on the working hypothesis that sonorant cues are detected independently in different parts of the frequency spectrum. This hypothesis provides a number of points of departure for our work. In the front-end, for example, it does not accommodate signal processing based on measurements of cepstra or wideband autocorrelation. Likewise, in the back-end, it does not accommodate “black box” statistical methods that violate the assumption of independence. Our approach also differs from purely “expert-based” approaches to feature detection. Though informed by aspects of human speech perception, we make elaborate use of statistical methods to fit models consistent with our working hypothesis. In fact, our approach reflects a more general trend in ASR and other areas of artificial intelligence: the design of rich statistical models with highly structured dependencies incorporating prior knowledge (Jordan, 1999). In our view, these probabilistic graphical models provide a way to study feature detection that combines the benefits of prior knowledge and learning from examples. They enable researchers to bridge the divide between models based on expert engineering (Holmes, 1998) and those derived by automatic methods (Bendiksen & Steiglitz, 1990).

4.3. Open problems

Much more work is needed, and in many areas. First, the OR combination of [+sonorant] cues in different critical bands is too simplistic. The OR gate, triggered by periodic activity in any part of the frequency spectrum, is obviously not appropriate for periodic forms of noise. In general, there must occur a more sophisticated integration of information across critical bands. It is commonly believed that listeners use commonalities in pitch and amplitude modulations to group different parts of the frequency spectrum into auditory streams (Bregman, 1994). In our setup, this suggests that cues for [+sonorant] should only accumulate across bands if these bands belong to the same stream. The ability to handle periodic interference, as arises from overlapping speakers, thus requires some sort of streaming mechanism.

A second (and related) assumption that needs to be relaxed is the complete independence of signal processing in different critical bands. Phenomena such as comodulation masking release (CMR) show that listeners correlate temporal modulation activity across critical bands (Hall, Haggard & Fernandes, 1984). The CMR effect refers to the striking observation that a pure tone in narrowband noise is more easily detected when flanking bands of masking noise are added with the same envelope modulations. It seems likely that a similar mechanism—comparing envelope fluctuations across critical bands—helps listeners to detect periodicity cues in noisy speech.

A third direction for research is to extend the model in this paper to other phonetic features. Clearly, not all features will be amenable to this approach. The approach seems most reasonable for those features, such as voicing and nasality, that are also extremely robust to noise and filtering (Miller & Nicely, 1955; Wang & Bilger, 1973). Other features will require different strategies for signal processing and integrating information across critical bands: a robust voicing detector, for example, might compute narrowband estimates of the voice onset time (Niyogi & Ramesh, 1998). Nevertheless, the network in Figure 2 can serve as a useful starting point for learning from examples of wideband speech. The main idea is that the hidden variables in these networks should encode a distributed representation of speech-related cues in different parts of the frequency spectrum. It would also be worthwhile to investigate other binary distinctions—such as rising/falling pitch, male/female speaker classification, or foreground/background identification—that can be made in narrow bands of speech.

A fourth and final challenge is to fold all these ideas into a speech recognizer. A simplistic way to do this is to compute phoneme probabilities by combining wideband measures of [+/−sonorant] and other phonetic features. The narrowband measures of [+sonorant] in Figure 5, however, convey much more information than the wideband measures in Figures 3 and 4. Ideally, uncertainties in narrowband feature detection (as reflected by the probabilities of hidden variables) should also be propagated so that they can be resolved by higher level considerations, such as linguistic context. Finally, it must be recognized that features do not turn on and off in perfect synchrony at phoneme boundaries. Traditional HMMs are not designed to handle this type of parallel asynchronous input, corresponding to the activation and deactivation of phonetic features (or partial cues) in different parts of the frequency spectrum. Recently, however, a number of researchers have investigated models that address these issues (Hopfield, Brody & Roweis, 1998; Mirghafori & Morgan, 1999). All these ideas need to be further developed.

We are grateful to H. Hwang (M.I.T.) for sharing the results of related experiments. We would also like to thank the editor, M. Ostendorf (U. Washington), and the anonymous reviewers for many valuable comments.

References

Ali, A. M. A., Van der Spiegel, J., Muller, P., Haentjens, G. & Berman, J. (1999). An acoustic-phonetic feature-based system for automatic phoneme recognition in continuous speech. *Proceedings of the International Symposium on Circuits and Systems*, Hong Kong, May 1999, pp. 118–121. IEEE.

Allen, J. B. (1994). How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing*, **2**, 567–577.

Bendiksen, A. & Steiglitz, K. (1990). Neural networks for voiced/unvoiced speech classification. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing '90*, Albuquerque, pp. 521–524.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, U.K.

Bourlard, H. & Dupont, S. (1996). A new ASR approach based on independent processing and recombination of partial frequency bands. *Proceedings of the International Conference on Spoken Language Processing '96*, Philadelphia, pp. 422–425.

Bourlard, H. & Dupont, S. (1997). Sub-band-based speech recognition. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing '97*, Munich, pp. 1251–1254.

Bregman, A. S. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, Cambridge, MA, U.S.A.

Chomsky, N. & Halle, M. (1968). *The Sound Pattern of English*, MIT Press, Cambridge, MA, U.S.A.

Clark, J. & Yallop, C. (1995). *An Introduction to Phonetics and Phonology*, Blackwell Publishing Ltd, Oxford, U.K.

Cooke, M., Green, P., Josifovski, L. & Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, in press.

Dempster, A., Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, **39**, 1–38.

Deng, L. & Sun, D. (1994). A statistical approach to ASR using atomic units constructed from overlapping articulatory features. *Journal of the Acoustical Society of America*, **95**, 2702–2719.

Erler, K. & Freeman, G. H. (1996). An HMM-based speech recognizer using overlapping articulatory features. *Journal of the Acoustical Society of America*, **96**, 2500–2513.

Espy-Wilson, C. (1994). A feature-based semivowel recognition system. *Journal of the Acoustical Society of America*, **96**, 65–72.

Fletcher, H. (1953). *Speech and Hearing in Communication*, Van Nostrand, New York.

Garofolo, J. S. (1988). *Getting Started With the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database*, National Institute of Standards and Technology (NIST), Gaithersburgh, MD.

Hall, J. W., Haggard, M. P. & Fernandes, M. A. (1984). Detection in noise by spectro-temporal pattern analysis. *Journal of the Acoustical Society of America*, **76**, 50–56.

Hartmann, W. A. (1997). *Signals, Sound, and Sensation*, Springer-Verlag, New York.

Hermansky, H., Pavel, M. & Tibrewala, S. (1996). Towards ASR on partially corrupted speech. *Proceedings of the International Conference on Spoken Language Processing '96*, Philadelphia, pp. 462–465.

Hess, W. (1983). *Pitch Determination of Speech Signals: Algorithms and Devices*, Springer-Verlag, New York, NY, U.S.A.

Holmes, J. N. (1998). Robust measurement of fundamental frequency and degree of voicing. *Proceedings of the International Conference on Speech and Language Processing '98*, Seattle, pp. 1007–1010.

Hopfield, J., Brody, C. & Roweis, S. (1998). Computing with action potentials. In *Advances in Neural Information Processing Systems*, (Jordan, M., Kearns, M. and Solla, S., eds), volume 10, pp. 166–172. MIT Press, Cambridge, MA, U.S.A.

Jankowski, C., Kalyanswamy, A., Basson, S. & Spitz, J. (1990). NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing '90*, Albuquerque, pp. 109–112.

Jordan, M, ed. (1999). *Learning in Graphical Models*, MIT Press, Cambridge, MA, U.S.A.

King, S. & Taylor, P. (2000). Detection of phonological features in continuous speech using neural networks. *Computer Speech and Language*, **14**, 333–353.

Kirchhoff, K. (1999). *Robust speech recognition using articulatory information*. PhD Thesis, University of Bielefeld, Germany.

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication*, **22**, 1–15.

Liu, S. (1996). Landmark detection for distinctive feature-based speech recognition. *Journal of the Acoustical Society of America*, **100**, 3417–3430.

Maron, O. & Lozano-Perez, T. (1998). A framework for multiple-instance learning. In *Advances in Neural Information Processing Systems*, (Jordan, M., Kearns, M. and Solla, S., eds), volume 10, pp. 570–576. MIT Press, Cambridge, MA, U.S.A.

Miller, G. A. & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, **27**, 338–352.

Ming, J. & Smith, F. J. (1999). Union: a new approach for combining sub-band observations for noisy speech recognition. *Proceedings of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, pp. 175–178.

Mirghafori, N. (1998). *A multi-band approach to automatic speech recognition*. PhD Dissertation, University of California Berkeley, U.S.A.

Mirghafori, N. & Morgan, N. (1999). Sooner or later: exploring asynchrony in multi-band speech recognition. *Proceedings of Eurospeech-99*, Budapest, pp. 595–598.

Moore, B. C. J. (1997). *An Introduction to the Psychology of Hearing*, Academic Press, San Diego, CA.

Moore, B. C. J. & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, **74**, 750–753.

Niyogi, P., Burges, C. & Ramesh, P. (1999). Distinctive feature detection using support vector machines. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing '99*, Phoenix, pp. 425–428.

Niyogi, P. & Ramesh, P. (1998). Incorporating voice onset time to improve letter recognition accuracies. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing '98*, Seattle, pp. 721–724.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo, CA, U.S.A.

Rabiner, L. R. & Juang, B. H. (1993). *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, U.S.A.

Slaney, M. An efficient implementation of the Patterson-Holdsworth auditory filterbank. Apple Computer Technical Report 35, Cupertino, CA.

Slaney, M. & Lyon, R. F. (1990). A perceptual pitch detector. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing '90*, Albuquerque, pp. 357–360.

Smith, L. A neurally motivated technique for voicing detection and F0 estimation for speech. CCCN Technical Report 22, University of Sterling, Sterling, Scotland.

Smith, L. (1997). A noise-robust auditory modelling front end for voiced speech. In *Artificial Neural Networks—ICANN-97*, Lecture Notes in Computer Science, volume 1327, (Gerstner, W., Germond, A., Hasler, M. and Nicoud, J.-D., eds), pp. 97–102. Springer-Verlag, Heidelberg, Germany.

Stevens, K. N. (1999). *Acoustic Phonetics*, MIT Press, Cambridge, MA, U.S.A.

Tibrewala, S. & Hermansky, H. (1997). Sub-band based recognition of noisy speech. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing '97*, Munich, pp. 1255–1258.

Varga, A., Steeneken, H. J. M., Tomlinson, M. & Jones, D. The noisex-92 study on the effect of additive noise on automatic speech recognition. Technical Report, DRA Speech Research Unit.

Wang, M. D. & Bilger, R. C. (1973). Consonant confusions in noise: a study of perceptual features. *Journal of the Acoustical Society of America*, **54**, 1248–1266.

(Received 28 February 2000 and accepted for publication 12 February 2001)

Appendix A: Front-end

Our front-end consists of a bank of 24 gammatone filters with overlapping passbands (Slaney, 1993). The widths of the passbands are matched to equivalent rectangular bandwidth (ERB) estimates of auditory filters (Moore & Glasberg, 1983), and the center frequencies are equally spaced on an ERB scale between 225 Hz and 4000 Hz. The outputs of these filters are half-wave rectified and squared. These nonlinearities are purposeful forms of intermodulation distortion (Hartmann, 1997). Specifically, in bands spanning two or more adjacent harmonics, these operations create and concentrate energy at the fundamental frequency (Smith, 1996). After these nonlinearities, the channels are bandlimited to 50–300 Hz and downsampled to speed up subsequent processing. The frequencies 50–300 Hz were chosen to represent the extremes in pitch of male and female adult speakers.

Measurements are made by blocking each channel into contiguous, nonoverlapping 16 ms frames, with a frame shift of 16 ms. Six measurements per frame are recorded for each channel. The first two measurements are running estimates of SNR, computed by dividing each frame's energy by the minimum energy of neighboring frames spanning 200 ms and 400 ms of speech, respectively. A small positive offset is added to the denominator in this calculation to accomodate regions of silence. If this ratio is greater than unity, the SNR is recorded as its logarithm; otherwise, it is simply recorded as 0 dB.

The autocovariance function for lags between 3.3 and 20 ms (50 and 300 Hz) is also computed in each channel, normalized by the value at zero lag plus a small positive offset. The normalization is used to compress the dynamic range of the speech signal. The autocovariances are computed over 64 ms windows centered on the same sample as the shorter 16 ms windows used to estimate the SNR. The longer windows are required to span multiple pitch periods of deep male voices. The maximum and minimum values of the autocovariance function are recorded as the third and fourth measurements in each frame, while the average values of peaks and valleys are recorded as the fifth and sixth measurements. In addition, the signs of the fourth and sixth measurements are flipped so that all the measurements are positively correlated with the amount of periodicity in the waveform.

Finally, the six measurements of SNR and autocovariance are thresholded to reduce their variance in noise. Threshold values are computed from the means plus one standard deviation of the same measurements for identically processed bands of white noise. Measurements greater than these threshold values are replaced by the amounts in excess; measurements less than these threshold values are replaced by zero. The purpose of these thresholding operations is to preserve only those sonorant cues unlikely to have been generated by noise.

Appendix B: EM algorithm

The EM algorithm for the network in Figure 2 attempts to minimize the cross entropy error function, given by Equation (7). Combining Equations (2) and (3), we can rewrite this error function as:

$$\mathcal{E} = -\sum_t \left\{ z_t \log\left(1 - \prod_i \left[1 - \prod_j p_{ij}^t\right]\right) + (1 - z_t)\log\left(\prod_i \left[1 - \prod_j p_{ij}^t\right]\right)\right\}, \tag{B1}$$

where the sum is over frames of speech, indexed by t. Note that the error function depends in a complicated way on the probabilities p_{ij}^t (and hence the parameters $\boldsymbol{\theta}_{ij}$) computed by the logistic regressions in the bottom layer of the network.

The EM algorithm provides an iterative procedure for minimizing the error function in Equation (B1), with guarantees of monotonic convergence. The algorithm consists of two alternating steps, an E-step and an M-step. The E-step for the model computes the posterior probabilities of the hidden variables, conditioned on the labels provided by the phonetic alignment. The calculations here are different for [–sonorant] and [+sonorant] frames of speech. For [–sonorant] frames, we have:

$$\Pr[X_{ij} = 1|Z = 0, \boldsymbol{M}] = p_{ij}\left[\frac{1 - \prod_{k\neq j} p_{ik}}{1 - \prod_l p_{il}}\right]. \tag{B2}$$

To make this equation easier to read, we have dropped the superscript indexing the frame number, which strictly speaking should be added to each variable in the equation (e.g. X_{ij}^t, Z^t, etc.). The posterior probabilities for [+sonorant] frames are given by:

$$\Pr[X_{ij} = 1|Z = 1, \boldsymbol{M}] = p_{ij}\left[\frac{1 - \prod_{k\neq j} p_{ik}}{1 - \prod_l p_{il}}\right]\left[1 - \frac{\prod_n p_{in}}{1 - \prod_m(1 - \prod_l p_{ml})}\right] + \frac{\prod_k p_{ik}}{1 - \prod_m(1 - \prod_l p_{ml})}. \tag{B3}$$

These posterior probabilities are derived by applying Bayes rule to the left-hand sides of Equations (B2) and (B3), marginalizing the hidden variable Y_i, and making repeated use of Equations (5) and (6). The computations remain tractable due to the conditional independencies of the underlying Bayesian network (Pearl, 1988).

The M-step of the EM algorithm updates the parameters in each logistic regression. Specifically, it prescribes how to choose updated parameter estimates, $\tilde{\boldsymbol{\theta}}_{ij}$, to replace the current ones, $\boldsymbol{\theta}_{ij}$. Let $q_{ij}^t = \Pr_\theta[X_{ij}^t = 1|Z^t = z^t, \boldsymbol{M}^t]$ denote the posterior probabilities computed from Equations (B2) and (B3), using the current parameter estimates, $\boldsymbol{\theta}_{ij}$. Likewise, let $\tilde{p}_{ij}^t = \Pr_{\tilde{\theta}}[X_{ij}^t = 1|\boldsymbol{M}^t]$ denote the prior probabilities computed from Equation (1), using the updated parameter estimates, $\tilde{\boldsymbol{\theta}}_{ij}$. The M-step consists of replacing $\boldsymbol{\theta}_{ij}$ by $\tilde{\boldsymbol{\theta}}_{ij}$, where:

$$\tilde{\boldsymbol{\theta}}_{ij} = \arg\max_{\tilde{\boldsymbol{\theta}}_{ij}}\left\{\sum_t [q_{ij}^t \log \tilde{p}_{ij}^t + (1 - q_{ij}^t)\log(1 - \tilde{p}_{ij}^t)]\right\}. \tag{B4}$$

Note that this procedure decouples the problem of parameter estimation into several independent weighted logistic regressions. The terms in Equation (B4) define a convex function of $\tilde{\boldsymbol{\theta}}_{ij}$, so that these maximizations can be performed by Newton's method or (in rare cases of instability) by gradient ascent. The power of the EM algorithm is that it replaces the seemingly intractable cross entropy error function in Equation (B1) by the simpler ones in Equation (B4).