Draft: please do not distribute.

LEARNING FROM EXAMPLES IN CRITICAL BANDS OF SPEECH

Lawrence K. Saul *Mazin G. Rahim* *Jont B. Allen*
AT&T Labs - Research, 180 Park Ave, Florham Park, NJ 07932

ABSTRACT

We propose to mimic auditory strategies for recognizing noisy or distorted speech. Motivated by psychoacoustic data on human performance, we begin by studying how to detect the phonetic feature of voicing. We describe an automatic method for voicing detection starting from a cochlear filterbank. Our method has three stages: first, speech is filtered into critical bands and enhanced by nonlinearities; second, a degree of voicing is computed in each band based on its periodicity and signal-to-noise ratio; third, signals from different bands are combined to make a global decision. These stages are formulated as components of a probabilistic graphical model, represented by a multilayer Bayesian network. We train the model from a corpus of phonetically transcribed speech, then evaluate its robustness to noise and corrupting influences. Implications for automatic speech recognition are discussed.

1. INTRODUCTION

Deprived of linguistic context, visual input, and binaural cues, humans maintain a robust capacity to recognize speech in poor listening conditions. This was first demonstrated by the pioneering experiments of Fletcher[?], in which subjects were asked to recognize nonsense syllables in noisy or distorted speech, without the benefit of auxiliary cues. These experiments, and many performed since then[?], support the idea that the robustness of human speech recognition is due in large part to the information processing at the auditory periphery—i.e., between the cochlea and the auditory cortex.

The resistance of speech to corrupting influences begins with the cochlear filtering into critical bands[?]. Critical bands are to the ear what pixels are to the eye—a mechanism for breaking down complicated problems in scene analysis[?]. Experiments on the articulation index have shown that cues for recognizing speech are distributed in a redundant manner across the frequency spectrum. Put another way, speech is a "spread spectrum" modulation scheme. Listeners recognize noisy or distorted speech by integrating evidence from different parts of the spectrum, ignoring bands with poor signal-to-noise ratios (SNRs), and capitalizing on bands with distinctive acoustic signatures[?].

Auditory strategies of information processing provide many insights into the problems of automatic speech recognition (ASR). Today's speech recognizers are especially sensitive to background noise and distortion. 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[?]. Viewed as a long-term strategy for matching human performance, this program for ASR has inherent limitations. The HMM back-end does not support the multiband processing used by the auditory system. Also, the cepstral front-end suppresses the periodic signature of the speaker's pitch, blurring the basic distinction between voiced and unvoiced speech and ignoring the phonetic information carried by this "voicing bit".

We feel that the mimicking of auditory strategies presents a compelling alternative approach to ASR. We do not dispute the power of statistical methods, combined with the availability of large amounts of data, to compensate for our incomplete understanding of auditory processes[?]. On the other hand, a distinction needs to be made between incomplete understanding and willful disregard. Certain fundamentals—such as filtering into critical bands—are so well established that it seems more logical to ask what types of statistical models incorporate these fundamentals, rather than what types of signal processing work best with HMMs.

In this paper, we propose a statistical model for detecting the phonetic feature of voicing—that is, for recognizing speech sounds generated by vibration of the vocal cords. As part of a long-term strategy for ASR, we view the ability to detect voiced speech in difficult listening conditions as a fundamental prerequisite for making finer distinctions (such as between phones, syllables, or words) with the same degree of robustness as human listeners. This view is based on psychoacoustic evidence that voicing determinations are made early in the speech chain, with greater reliability than those of other phonetic features[?].

Important challenges for ASR are to solve two basic problems: first, to identify voiced segments in noisy, band-limited speech, making few assumptions about the noise, channel, or bandwidth; second, to determine the voicing profile as a function of frequency—that is to detect which regions of the frequency spectrum provide evidence of voicing, and which do not. In our view, existing methods for voicing detection are inadequate. On one hand, there are heuristic algorithms based on auditory processes; these algorithms, relying mainly on expert knowledge and manual parameter-tuning, provide no guarantee of optimality on large data sets. On the other hand, there are general purpose methods from statistical pattern recognition; these methods, ignoring the role of critical bands, do not generalize well to noisy, bandlimited speech.

Our approach to voicing detection combines the strengths of auditory and statistical models. These strengths are the ideas of multiband processing and learning from examples. Our front-end consists of a cochlear filterbank followed by half-wave rectification and other nonlinearities. Narrow-band measurements of SNR and periodicity are then fed to a statistical model, represented by a multilayer Bayesian network, whose binary hidden variables indicate voiced excitation in different parts of the frequency spectrum. The structure of the model formalizes a logical hypothesis about how information is combined across critical bands. The model is discriminatively trained to maximize the probability of correct voiced-unvoiced classification in clean speech. Results of voicing detection in clean and corrupted speech demonstrate the robustness of our approach.

2. VOICING

Psychoacoustic data show that voicing is one of the first bits of information extracted from the speech signal, as well as one of the least likely to be corrupted. Miller and Nicely[?] measured the intelligibility of nonsense syllables in noisy and bandlimited speech. Tabulating confusion matrices, they found that over a wide range of acoustic conditions, listeners were much less likely to confuse voiced consonants for unvoiced ones, and vice versa. These results imply that the phonetic feature of voicing is detected early in the speech chain, prior to the more general recognition of phones, syllables, or words. They also imply that voicing is detected more robustly than other phonetic features: most confusions in noisy or bandlimited speech do not involve voicing.

In light of these results, it is somewhat ironic that most ASR front-ends begin by systematically erasing all signs of voicing[1] from the waveform. While possible to recognize devoiced speech (e.g., whispering), such a strategy seems unlikely to match the robustness of human listeners. Whispered speech is not as resistant to corrupting influences as normal speech. In fact, its typical purpose is to prevent all but the most ideally situated listeners from understanding what is being said.

Based on auditory strategies, we believe that the "voicing bit" of phonetic information should be the first priority of signal processing in ASR, not the first casualty. Speech recognition is an analog-to-discrete process in which acoustic waveforms are converted to discrete sequences of symbols. The detection of phonetic features[?], such as voicing, marks the perceptual boundary at which analog information is converted to discrete form. Errors at this stage can be (and are) corrected by higher-order linguistic processes. But such correction is limited, time-consuming, and mentally taxing. Semantic and syntactic processes cannot correct widespread errors in phonetic feature detection; the time scales here are simply not commensurate.

Successful ASR requires a balanced integration of bottom-up and top-down strategies. Many researchers have proposed a bottom-up component based on the detection of phonetic features, starting with voicing. The so-called "acoustic-phonetic" approach advocates a divide-and-conquer strategy for extracting phonetic bits of information from the speech signal. This approach has been criticized for relying on questionable assessments of phonetic features, and for not exploiting statistical methods that provide some guarantee of optimality on large data sets[?]. Mindful of these criticisms, this work focuses on a phonetic feature of widely recognized importance and makes a rather purposeful use of statistical methods.

The acoustic-phonetic approach is premised on the ability to detect phonetic features in noisy or distorted speech and to label which parts of the frequency spectrum constitute evidence for positive identifications. We attribute previous difficulties with this approach to the challenge of detecting phonetic features in narrow bands of speech. Coordinated efforts in signal processing and statistical modeling are required to build phonetic feature detectors. This leads us to consider the information processing strategies of the peripheral auditory system.

3. CRITICAL BANDS

In voiced speech, energy is concentrated at harmonics, or equally spaced multiples of the speaker's pitch. Periodicity at the fundamental frequency can be detected in critical bands containing sufficient energy at two or more adjacent harmonics. When certain parts of the frequency spectrum are corrupted by noise or filtering, voicing determinations are made from surviving critical bands with high SNR. Our strategy for voicing detection is based on this picture of auditory processes. We focus on two acoustic correlates of voicing: the periodicity established by the speaker's pitch, and changes in the signal-to-noise ratio (SNR).

Our front-end consists of a bank of 24 bandpass filters with overlapping passbands. The center frequencies are equally spaced on a logarithmic scale between 250 Hz and 3600 Hz, and the width of each passband is matched to empirical estimates of cochlear bandwidth[?]. Cochlear filters are modeled (crudely) by 2rd order type-I Chebyshev filters.

The outputs of these filters are half-wave rectified and squared. The half-wave rectification mimics the transduction of neural firing patterns by the inner hair cells; it is known that neural excitation occurs only for one direction of movement of the basilar membrane[?]. The squaring nonlinearity is a purposeful form of intermodulation distortion. In voiced bands spanning two or more adjacent harmonics, the squaring operation creates energy at the fundamental frequency of the speaker's pitch. After these nonlinearities, the channels are bandlimited to 50–300 Hz and downsampled to speed up subsequent processing.

Measurements are made by blocking each channel into 64 ms frames with a frame shift of 10 msec. Five measurements per frame are recorded in each channel. The first measurement is a running estimate of SNR, computed by dividing each frame's energy by the minimum energy of neighboring frames spanning 400 ms of speech. 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 recorded as 0 dB. An autocovariance for lags between 50 Hz and 300 Hz is also computed for each frame, 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 maximum and minimum values of this autocovariance are recorded as the second and third measurements, while the average values of peaks and valleys are recorded as the fourth and fifth measurements. These values provide a measure of the periodicity at frequencies within the normal range of pitch for an adult speaker.

To summarize, our front end transforms the speech waveform into an array of baseband signals with energy concentrated between 50–300 Hz. Parallel measurements are made on sliding windows of speech in each band. The first measurement is an estimate of SNR; the remaining four are autocovariance statistics. There are 24 channels, resulting in a total of 120 measurements per 64 ms window of speech. One hundred such frames are processed per second. These

[1] Typically, the first stage of signal processing is to compute smooth estimates of the short-time magnitude spectra; phase information is discarded, and the periodic signature of the speaker's pitch is completely suppressed.

frames are fed one at a time (or in small groups) to the statistical model described in the next section.

4. STATISTICAL MODEL

Many statistical models for voicing detection have been proposed in the literature[?]. Generally speaking, models trained in clean environments do not generalize well to noisy, bandlimited speech. One strategy to improve generalization is to incorporate noisy speech into the training procedure. By itself, however, this strategy represents a flailing, shortsighted approach to the problem of robustness. It fails to appreciate the number of degrees of freedom in noisy, bandlimited speech. One cannot hope to sample the universe of acoustic possibilities in any reasonable training procedure. We feel that the only viable long-term strategy for robustness is to incorporate prior knowledge about the speech signal[?].

Voicing detection in our model occurs in two stages. In the first stage, an independent assessment of voicing is made in each critical band. For a critical band to be labeled as voiced, its measurements of SNR and autocovariance must pass a number of statistical tests. In the second stage, information is combined across critical bands. In this work, we follow a simple rule: for the wideband speech to be labeled as voiced, it must be the case that voicing is detected in one or more critical bands.

Figure ?? depicts our model as a multilayer Bayesian network[?]. The nodes in this network represent random variables, while the links represent conditional dependencies. The bottom-up flow of the network describes the sequence of operations used to determine if a frame of speech is voiced.

The nodes in the bottom layer represent the measurements of SNR and autocovariance in each critical band. These variables are always instantiated—that is, determined by measurements of the waveform—whereas the other nodes in the network represent binary random variables with genuine uncertainty. The nodes in the bottom layer are shaded to emphasize this distinction.

The nodes in the next layer represent the outcomes of statistical tests on the measurements of SNR and autocovariance. We use $\boldsymbol{M}_i$ to denote the vector of measurements in the ith critical band, and X_{ij} to specify the outcome of the jth test in this band. The conditional probability that the jth test in the ith band is satisfied is given by:

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

where $\sigma(z) = [1+e^{-z}]^{-1}$ is the logistic function. The weights $\boldsymbol{\theta}_{ij}$ in this equation may be viewed as parameters in a logistic regression. Though not explicitly indicated, a bias term can be included in the logistic regression by adding an extra input to the measurement vector, $\boldsymbol{M}_i$.

The nodes in the next layer represent the positive or negative assessments of voicing 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 for voicing is made only if all the tests X_{ij} are positive. Thus, the conditional probability distribution for Y_i is given by:

$$\Pr[Y_i=1|\boldsymbol{M}_i] \;=\; \prod_j \Pr[X_{ij}=1|\boldsymbol{M}_i]. \qquad (2)$$

The nodes in this layer are depicted as AND gates to indicate the conjunction relating X_{ij} to Y_i. The AND gates, *triggered only by a consensus of positive inputs*, are designed to minimize false positives in noise.

Figure 1. Multilayer Bayesian network for voicing detection.

The top node in the network represents the overall assessment of voicing for the frame of wideband speech. We use Z to denote this binary random variable. As described earlier, the frame is labeled as voiced if and only if one or more critical bands are labeled as voiced. Thus, the conditional probability distribution for Z is given by:

$$\Pr[Z=1|\boldsymbol{M}] \;=\; 1-\prod_i\left(1-\Pr[Y_i=1|\boldsymbol{M}_i]\right). \qquad (3)$$

Here, we have used $\boldsymbol{M}=\{\boldsymbol{M}_1,\boldsymbol{M}_2,\ldots\}$ to denote the entire set of measurements. The top node in the network is depicted as an OR gate to indicate the disjunction relating Y_i to Z. The OR gate, *silenced only by a consensus of negative inputs*, is designed to minimize false negatives—i.e., failures to detect voicing in noisy or bandlimited speech.

The quantitative predictions of this model are determined by the weights $\boldsymbol{\theta}_{ij}$ in the bottom layer of the network. These weights are estimated from a corpus of phonetically transcribed speech. The training data for this procedure consists of frames of wideband speech, labeled as voiced or unvoiced based on the phonetic transcription. The parameters are chosen to maximize the likelihood that the model's predictions (i.e., $\Pr[Z|\boldsymbol{M}]$) match the labels generated by the phonetic transcription. The voicing determinations in individual critical bands (i.e., $\Pr[Y_i|\boldsymbol{M}_i]$) are modeled as hidden variables. Monotonic convergence to a local maximum in the likelihood is guaranteed by the EM algorithm[?], a general iterative procedure for parameter estimation in statistical models with hidden variables. Further details of this algorithm will be provided in a longer article.

Two simple extensions to the above model are useful for producing smooth voicing 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 statistical tests under each AND gate. These extensions do not complicate the inference procedures in any way.

5. EXPERIMENTS

The multiband model was trained on speech from the first dialect region of the TIMIT speech corpus. Voicing labels were generated from the phonetic transcription, with the initial assignments based on linguistic conventions[?]. An iterative procedure was then used to optimize these assignments, incorporating the effects of left and right context. The goal of this procedure was to determine the "bit" of phonetic information most readily available from critical band measurements of SNR and periodicity. The final assignments were close, but not identical, to the conventional categorization of voiced and unvoiced phones.

We also investigated how well current acoustic models in ASR capture the same bit of phonetic information. Two Gaussian mixture models (GMMs)—one for voiced speech, one for unvoiced speech—were trained from windowed 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. Each mixture model had 16 components with

Figure 2. Frame error rates for voicing detection in ten different environments.

diagonal covariance matrices. The number of mixture components was selected to optimize the voicing detection in a matched testing condition.

The multiband model and GMMs were trained exclusively on clean speech. Training data consisted of 380 sentences, or 112586 frames of speech (54% voiced, 46% unvoiced). We then evaluated the robustness of these models to noise and corrupting influences. This was done by measuring the frame error rate of their voicing determinations in corrupted speech. Testing data consisted of 110 sentences from different speakers in the same dialect region. The models were evaluated in ten testing conditions: (1) clean speech (matched conditions), (2) telephone distortion (nTIMIT), (3) 0 dB white noise, (4-7) 0 dB noise from 0-1 kHz, 1-2 kHz, 2-3 kHz, and 3-4 kHz, and (8-10) speech bandlimited to 0-1 kHz, 1-2 kHz, and 2-3 kHz.

Figure ?? shows the results of these experiments. The multiband model is significantly more robust. Note that severe noise and filtering lead to false negatives in the multiband model (as one might expect even from humans), but to random errors in the GMMs. These results shed light on the fragility of current recognizers. Similar results have been obtained for speech contaminated by babble noise, white noise, pink noise, and factory noise at SNRs from 30 dB to -5 dB. These results will be presented in a longer article.

6. SUMMARY

The multiband processing in our model makes it robust to many types of noise and distortion. To avoid "misses", or failures to detect voicing, the model exploits the idea that evidence for voicing is distributed across the frequency spectrum. Likewise, to avoid false positives, the model looks for multiply consistent measurements of periodicity and SNR. These measurements are derived from cochlear filters and purposeful nonlinearities, as opposed to spectral energies and cepstra, which have little predictive value in corrupted speech.

The architecture of our model is specifically tailored to the problem of multiband voicing detection. This gives it a unique advantage over generic approaches, such as decision trees, fully connected neural networks, or mixture models. Notably, our model learns to provide not only an overall estimate of voicing, but also a profile by frequency (i.e., the voicing determinations in individual critical bands, $\Pr[Y_i|M_i]$). Beyond the problem of voicing detection, our approach illustrates a more general trend in the field of artificial intelligence—the design of rich statistical models with highly structured dependencies incorporating expert knowledge. Probabilistic graphical models are at the core of a renaissance in the fields of artificial intelligence and neural networks. The potential to combine these models with our current understanding of auditory processes has not been fully realized. We feel it is worth revisiting the acoustic-phonetic paradigm for ASR in this framework.

Other researchers have also studied multiband processing[?] and Bayesian networks for ASR[?]. Our goal has been to combine these ideas with insights from auditory processes and psychoacoustics. The present work eschews the trappings of cepstra, HMMs, and word error rates and focuses on the fundamental problem of voicing detection. This may seem like a step backwards from the frontier of large vocabulary ASR. However, we believe it is a necessary step to overcome the fragility of current methods.

REFERENCES

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

[2] A. S. Bregman (1994). *Auditory Scene Analysis: the Perceptual Organization of Sound.* MIT Press: Cambridge, MA.

[3] K. N. Stevens (1986). Models of phonetic recognition II: A feature-based model of speech recognition. In P. Mermelstein, ed. *Proc. Montreal Satellite Symposium on Speech Recognition.* 12th International Congress of Acoustics: Montreal.

[4] A. Dempster, N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B* 39:1–38.

[5] H. Fletcher (1953). *Speech and Hearing in Communication.* Krieger: New York, NY.

[6] W. Hess (1983). *Pitch Determination of Speech Signals: Algorithms and Devices.* Springer Verlag: Berlin.

[7] J. Makhoul and R. Schwartz (1985). Ignorance modeling. In J. S. Perkell and D. H. Klatt, eds. *Variability and Invariance in Speech Processes.* Lawrence Erlbaum Assoc: Hillsdale, NJ.

[8] G. A. Miller and P. E. Nicely (1955). An analysis of perceptual confusions among some English consonants. *Acoust. Soc. Am. J.* 27(2):338–352.

[9] B. C. J. Moore (1997). *An Introduction to the Psychology of Hearing.* Academic Press.

[10] B. C. J. Moore and B. R. Glasberg (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J. Acoust. Soc. Am. 74:750–753.

[11] J. Pearl (1988). *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann: San Mateo, CA.

[12] L. R. Rabiner and B. H. Juang (1993). *Fundamentals of Speech Recognition.* Prentice Hall: Englewood Cliffs, NJ.

[13] S. Tibrewala and H. Hermansky (1997). Sub-band based recognition of noisy speech. In Proc. ICASSP-97, 11:1255-1258.

[14] V. W. Zue (1995). The use of speech knowledge in automatic speech recognition. *Proc. IEEE*, 73(11): 1602–1615.

[15] G. Zweig and S. Russell (1998). Probabilistic modeling with Bayesian networks for ASR. In Proc. ICSLP-98.