EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

USING MISSING FEATURE THEORY TO ACTIVELY SELECT FEATURES FOR ROBUST SPEECH RECOGNITION WITH INTERRUPTIONS, FILTERING, AND NOISE*

Richard P. Lippmann and Beth A. Carlson
email: rpl@SST.LL.MIT.EDU
Room S4-121, Lincoln Laboratory MIT
244 Wood Street
Lexington, MA 02173-9108
USA

ABSTRACT

Speech recognizers trained with quiet wide-band speech degrade dramatically with high-pass, low-pass, and notch filtering, with noise, and with interruptions of the speech input. A new and simple approach to compensate for these degradations is presented which uses mel-filter-bank (MFB) magnitudes as input features and missing feature theory to dynamically modify the probability computations performed in Hidden Markov Model recognizers. When the identity of features missing due to filtering or masking is provided, recognition accuracy on a large talker-independent digit recognition task often rises from below 50% to above 95%. These promising results suggest future work to continuously estimate SNR's within MFB bands for dynamic adaptation of speech recognizers.

1. INTRODUCTION

Humans can recognize speech with normally occurring degradations caused by head shadow, room coloration, and environmental noise. We can also recognize speech reproduced with bandwidth variability and noise introduced by modern communications devices and speech with other unnatural distortions such as sharp high-pass and low-pass filtering [4], severe band-reject filtering [9], and extremely erratic linear frequency responses [5]. Machine recognition performance, however, often degrades dramatically with channel variability and noise, even when noise and channel compensation is applied [8].

Many current compensation techniques estimate noise and channel characteristics off-line, and then adapt internal recognition parameters to values that would have been produced by training in the degraded environment. These techniques are limited because they are computationally expensive, they have been applied primarily with static frequency response variability and noise, and they often require long isolated speech and/or noise samples. This paper introduces a simple alternative approach motivated by missing feature theory (e.g. [1],[2]) and the effortless and fast adaptation exhibited by humans.

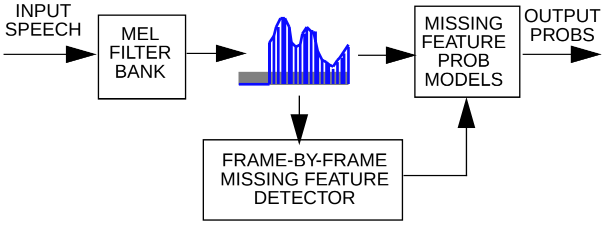

INPUT SPEECH
MEL FILTER BANK
MISSING FEATURE PROB MODELS
OUTPUT PROBS
FRAME-BY-FRAME MISSING FEATURE DETECTOR

FIGURE 1. Block diagram of a speech recognizer with missing-feature adaptation.

2. MISSING FEATURE THEORY

Figure 1 shows a block diagram of a speech recognizer which uses missing feature theory to improve recognition performance. The input speech waveform is processed using a mel-filter-bank (MFB) analysis [12] to produce a set of MFB log spectral magnitude features for every new speech input frame. The recognizer forms probabilistic models in HMM nodes of these features directly instead of converting them into a smaller number of cepstral features. This direct use of spectral magnitude features increases the number of input features and computational requirements, but greatly simplifies that task of compensating for missing features. During every new input frame, each spectral magnitude feature is examined by a missing feature detector to determine whether it represents true input speech or noise. Figure 1 illustrates a simple missing feature detector where every spectral magnitude which falls below a low threshold value (shown as a gray region) is declared missing. The example in Figure 1 shows an input spectrum which has been high-pass filtered where the first eight spectral magnitude features are below threshold and are labeled as

*This work was sponsored by DARPA. Opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the United States Air Force.

"missing." Once each spectral magnitude feature in a frame is labeled as missing or present, a computationally simple modification of probability models discards missing features and forms densities which would have been obtained by training without missing features.

Missing feature theory can be applied with filtering or interruptions and also when speech is masked by noise. With all these degradations, MFB features in spectral regions where speech is attenuated below the noise floor at the input to the recognizer are considered missing. Missing feature theory is simple to apply to high-performance Gaussian-mixture HMM recognizers which use diagonal covariance matrices. During training, the forward-backward algorithm estimates parameters in Gaussian mixture likelihood functions $p(\boldsymbol{x}|c)$ for each HMM node, where $\boldsymbol{x}$ represents the input MFB feature vector, and c represents the sound class for an HMM node. When a specific set of features $\boldsymbol{x}_p$ is present and the remaining features are missing, then likelihood functions in HMM nodes must be replaced by $p(\boldsymbol{x}_p|c)$. With diagonal covariance Gaussian mixture modeling, the original full likelihood function can be expressed as a weighted sum of the products of univariate Gaussian densities.

$$p(\boldsymbol{x}|c) = \sum_{j=1}^{M} w_j \prod_{i=1}^{D} N(m_{ij}, \sigma_{ij}^2)\,, \qquad (1)$$

where M represents the number of mixture components, w_j represents the mixture weight for mixture component j, D represents the number of input features, and N represents a univariate Gaussian distribution function for input feature x_i and mixture component j, with variance σ_{ij}^2 and mean m_{ij}. When some of the mixture components are missing, this can be expressed as

$$p(\boldsymbol{x}|c) = \sum_{j=1}^{M} w_j \prod_{\substack{i \\ present}} N(m_{ij}, \sigma_{ij}^2) \prod_{\substack{i \\ mis\sin g}} N(m_{ij}, \sigma_{ij}^2)\,. \qquad (2)$$

In this equation, each mixture component is expressed as a product of one-dimensional Gaussian components for the features that are present times a product of Gaussian components for the features that are missing. The modified likelihood $p(\boldsymbol{x}_p|c)$ required to recognize speech with missing features can be obtained by integrating the original likelihood function $p(\boldsymbol{x}|c)$ over the missing features, $p(\boldsymbol{x}_p|c) = \int p(\boldsymbol{x}|c) d\boldsymbol{x}_m$, where $\boldsymbol{x}_m$ represents all missing features. This integration simply eliminates the right hand product in Equation 2 because each term in that product integrates to unity. The desired modified likelihood is then given by

$$p(\boldsymbol{x}_p|c) = \sum_{j=1}^{M} w_j \prod_{\substack{i \\ present}} N(m_{ij}, \sigma_{ij}^2)\,. \qquad (3)$$

This function can be calculated by simply dropping terms corresponding to missing features from the original full likelihood computation. Adapting density estimation computations requires very little extra complexity and requires only minimal modification of existing speech recognition software once the identity of missing features is known. It is thus easy to insert this compensation into a recognizer and also to vary the compensation from one frame to another for time-varying filtering or noise. A similar approach to missing-feature adaptation can be applied to Parzen window density estimation and the normalized radial-basis-function neural networks used in [1][3].

3. DIGIT RECOGNITION TASK

All digits spoken by male talkers in the Texas Instruments digit speech corpus [6] were used to evaluate the effectiveness of missing feature compensation. This talker-independent digit recognition task is difficult enough to determine the utility of the new approach, but small enough to allow evaluations using many test conditions. Training used data from 54 talkers (1188 tokens) and testing used data from the remaining 56 talkers (1232 tokens). Good performance for wide-band quiet speech sampled at a 20 kHz rate was obtained using an HMM recognizer with whole-word left-to-right word models with 8 nodes per word and two diagonal-covariance Gaussian mixtures per node. MFB log spectral magnitude features were computed every 10 msec using a 25 msec Hamming window. Thirty mel-spaced triangular filters with centers ranging from roughly 64 to 9,100 Hz were used to compute MFB log spectral magnitudes. Two reference recognizers were used to evaluate baseline performance without missing feature compensation. The first cepstral recognizer used 12 MFB cepstral input features and 12 delta cepstra features. The second MFB recognizer used 30 MFB log magnitude features and 30 delta MFB features. Reference recognizers were compared to a third recognizer which was identical to the MFB reference recognizer, but that used missing feature compensation. Although missing features could have been detected using a threshold detector as shown in Figure 1, the identity of missing features in all experiments was provided as a priori information and used to compensate for missing features during recognition. Results thus provide an upper bound on performance that could be obtained with perfect missing feature detection.

4. RESULTS WITH STATIC FILTERING

Initial experiments explored the effect of static, non-time-varying, linear filtering. In all experiments, recognizers were first trained using quiet wide-band speech, and then tested with filtered speech without retraining. Instead of filtering the input speech waveform, sharp filtering was approximated in the frequency domain by directly attenuating mel-filter-bank magnitudes. A constant value was subtracted from log mel-filter-bank magnitudes equivalent to an attenuation of roughly 40 dB, and resulting log values which fell below a fixed threshold were clipped at that threshold. The threshold

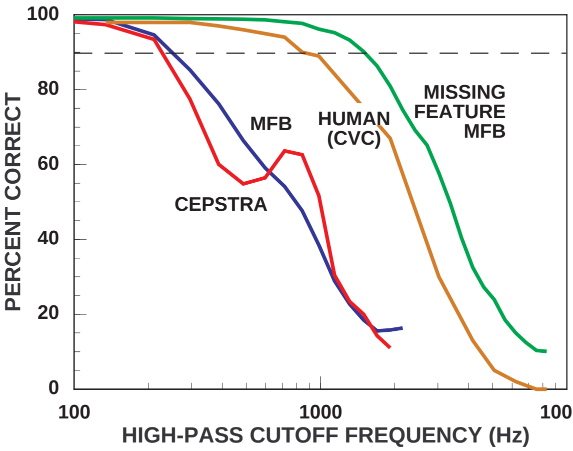


FIGURE 2. High-Pass Filtering Results.

was set to be roughly 10 dB below the lowest log magnitude levels observed with this speech corpus to simulate filtering followed by adding low-level noise which masks highly attenuated speech components.

The first experiments explored the effect of high-pass filtering. Figure 2 shows results obtained using the two reference recognizers and the MFB recognizer with missing feature adaptation. These results are compared to prior human results obtained on a much more difficult CVC nonsense syllable task [4]. Accuracy without filtering is roughly 99% correct for all machine recognizers. The accuracy for both reference recognizers degrades rapidly with only a small amount of filtering. Performance drops well below 90% correct with high-pass filtering when the cutoff frequency is above 300 Hz. Accuracy degrades rapidly for both reference recognizers with progressively more severe high-pass filtering until the recognizers are operating at chance levels with filtering at cutoff frequencies above roughly 2 kHz. Missing feature compensation dramatically improves performance. Accuracy with the missing feature MFB recognizer is above 90% correct even with extreme high-pass filtering that eliminates all energy below 1.5 kHz. Human and machine results with missing-feature compensation are similar in that performance doesn't degrade substantially with high-pass filtering up to a cutoff frequency of 1 kHz and performance drops off only gradually as the high-pass cutoff frequency is raised. Machine performance in Figure 2 appears better than human performance because the machine digit recognition task was much simpler than the human CVC nonsense syllable task (perplexity of 10 versus roughly 6,900), and the identity of missing features was known a priori.

Results of a second set of experiments which explored the effects of low-pass filtering are shown in Figure 3 along with prior human CVC nonsense syllable results [4]. The accuracy for both reference recognizers degrades rapidly down from roughly 99% correct with only a small amount of low-pass filtering. Performance drops well below 90% correct with a low pass cutoff below roughly 6 kHz and then degrades rapidly with progressively more severe low-pass filtering. Missing feature compensation again dramatically improves performance. Accuracy with the missing feature MFB recognizer is above 95% correct even with extreme low-pass filtering that eliminates all energy above 2 kHz. Human and machine results are again similar.

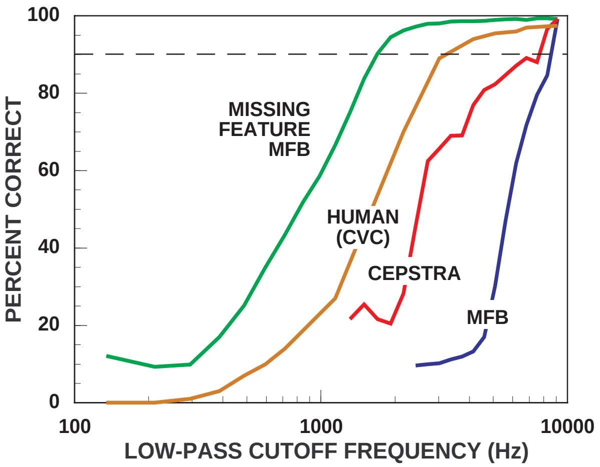


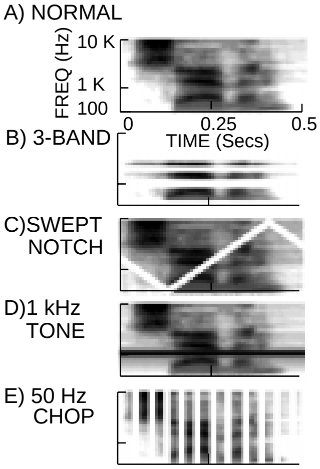
FIGURE 3. Low-Pass Filtering Results.

5. TIME-VARYING FILTERING AND NOISE

Further experiments explored the effect of four types of distortions which have been found to degrade human speech perception only slightly. Figure 4 shows spectrograms created using MFB log magnitude features for the word “seven,” spoken normally and with these distortions. It also shows recognition accuracy for two reference recognizers and the missing-feature MFB recognizer on the digit corpus. The upper row in the right-hand table of Figure 4 shows that the machine digit recognition accuracy for all recognizers is roughly 99% correct.

The effect of multiband filtering on the original “seven” spectrogram using three sharp 500 Hz passbands centered at 500, 1500, and 2500 Hz is shown in Figure 4B. This extremely erratic filtering provides a high human recognition accuracy of roughly 92% correct for words in meaningful sentences [5]. As can be seen in the right of Figure 4, recognition accuracy with the reference recognizers falls below 15% correct with this type of filtering, but missing-feature compensation improves performance to 93.3% correct.

The effect of a time-varying notch filter and of adding a 1 kHz pure tone on the original “seven” spectrogram is shown in Figure 4C and Figure 4D. Informal listening experiments suggest that these two distortions have little effect on intelligibility of words in meaningful sentences. The notch was created by reducing the levels of three adjacent MFB magnitudes in the front-end processing by 40 dB and by sweeping the center of the notch across the entire 10 KHz frequency range 5 times every second. Recognition accuracy for



REF. MFB	REF. CEP.	MISS MFB
98.8%	99.2%	98.8%
9.1%	14.3%	93.3%
14.2%	9.7%	98.9%
55.9%	11.8%	97.0%
9.1%	82.6%	96.6%

FIGURE 4. MFB Spectrograms for the word “seven” and recognition results under normal conditions and with four distortions.

both reference recognizers falls below 15% correct with this swept notch, but missing-feature compensation restores performance to a normal level of 98.9% correct. This performance is well within the 98.8±0.6% two binomial standard deviation range expected for this experiment. The effect of noise was evaluated by adding a 1 kHz pure tone at a signal-to-noise-ratio (SNR) of 10 dB to all speech waveforms, where the SNR was determined using the average RMS levels of all sampled and segmented speech waveform tokens. Missing feature compensation was applied by labeling all MFB magnitudes in the four bands nearest 1 KHz as noise. Recognition accuracy falls to 55.9% with the MFB recognizer and 11.8% with the cepstra recognizer. Missing-feature compensation restores performance to a near-normal level of 97.0% correct.

The final distortion illustrated in Figure 4E was to approximate the effect of interrupting the speech waveform every other 20 msecs. Past studies show that humans achieve accuracies of above 90% correct for words in sentences under this condition [10]. This condition was approximated by attenuating MFB features in every other pair of frames by 80 dB followed by thresholding to simulate low-level noise, as with filtering. Missing feature compensation was applied by setting the total likelihood score for missing frames to 1.0 and thus ignoring scores for missing frames. Performance drops to 9.1% with the MFB reference recognizer and to 82.6% with the cepstra recognizer. The reference cepstra recognize is more robust to this distortion because the shape of the flat attenuated input spectrum is not too extreme, while the low-level MFB log values are extreme outliers. Performance with missing feature compensation increases to a near-normal level of 96.6% correct.

6. SUMMARY AND DISCUSSION

This preliminary study demonstrates the simplicity and effectiveness of missing-feature compensation on a relatively simple digit recognition task when the identity of missing features is known a priori. Further studies are required to demonstrate the effectiveness of this approach with larger and more difficult speech corpora and when the outputs of speech/noise detectors are used to identify missing MFB features. Although developing highly accurate speech/noise detectors is a difficult task, filtering and frame-deletion results presented in this paper demonstrate that extreme accuracy is not required. A bias towards labeling MFB bands as missing has little effect on overall machine recognition accuracy, and many MFB features can be omitted before performance drops substantially.

7. REFERENCES

1. S. Ahmed and V. Tresp, “Some Solutions to the Missing Feature Problem in Vision”, In S. J. Hanson, J. D. Cowan, and C. L. Giles (Eds.), *Advances in Neural Information Processing Systems*, Volume 5, pp. 393-400, Morgan Kaufmann, San Mateo, 1993.
2. Cooke, M.P., Morris, A. & Green, P.D., “Recognizing Occluded Speech”, in Proceedings of the ESCA Tutorial and Research Workshop on *The Auditory Basis of Speech Perception*, Keele University, United Kingdom, 15-19 July, 1996, 297-300, Morgan Kaufmann, San Mateo, 1996.
3. E. Chang and R. Lippmann, “Improving Wordspotting Performance with Artificially Generated Data”, In Proceedings IEEE International Conference on Acoustics Speech and Signal Processing, 526-529, 1996.
4. N. R. French and J. C. Steinberg, “Factors Governing the Intelligibility of Speech Sounds,” Journal of the Acoustical Society of America, 19(1), 90-119, 1947.
5. K. D. Kryter, “Speech Bandwidth Compression through Spectrum Selection”, Journal of the Acoustical Society of America, 32(5), 547-556, 1960.
6. R. G. Leonard, “A Database for Speaker-Independent Digit Recognition”, in Proceedings IEEE International Conference on Acoustics Speech and Signal Processing, 42.11.1-41.11.4, 1984.
7. J. C. R. Licklider and I. Pollack, “Effects of Differentiation, Integration, and Infinite Peak Clipping upon the Intelligibility of Speech”, Journal of the Acoustical Society of America, 20(1), 42-51, 1948.
8. R. P. Lippmann, “Speech Recognition by Machines and Humans”, Journal of Speech Communication, In Press, 1997.
9. R. P. Lippmann, “Accurate Consonant Perception Without Mid-Frequency Speech Energy”, IEEE Transactions on Speech and Audio Processing, 4(1), 66-69, 1996.
10. Miller, G.A. and J.C.R. Licklider, “The Intelligibility of Interrupted Speech.” Journal of the Acoustical Society of America, 22(2), 167-173, 1950.
11. A. Varga and R. Moore, “Hidden Markov Model Decomposition of Speech and Noise”, In Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, 845-848, 1990.
12. S. J. Young, “A Review of Large-Vocabulary Continuous-Speech Recognition,” IEEE Signal Processing Magazine, 13(5), 45-57, 1996.