Proc. of Intl. Conf. Acoustic, Speech and Signal Proc., Salt Lake City, Utah, May 2001.

SOFT-FEATURE DECODING FOR SPEECH RECOGNITION OVER WIRELESS CHANNELS

Alexandros Potamianos and Vijitha Weerackody

Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, NJ 07974
E-mail: {potam,vijitha}@research.bell-labs.com

ABSTRACT

A distributed automatic speech recognition (ASR) system is considered where features of the speech signal are extracted at the wireless terminal and transmitted to a centralized ASR server. An unequal error protection scheme is used for the quantized ASR feature stream. At the receiver, coherent demodulation is performed and the probability of error for each bit is computed using the Max-Log MAP algorithm [5]. A 'soft-feature' decoding strategy is introduced at the ASR server that uses the marginal distribution of only the reliable features during likelihood computation. Alternatively, the confidence of each feature is computed from the bit error probabilities and each feature in the probability computation is weighted as a function of the feature confidence. The performance of the proposed soft-feature algorithms is evaluated over typical cellular wireless channels and it is shown to reduce ASR error rate by over 50% for certain channels at a small additional computational cost.

1. INTRODUCTION

Automatic speech recognition (ASR) over wireless networks is important for next generation wireless multimedia systems [3]. A variety of spoken dialogue systems exist today that utilize ASR technology, e.g., personal assistants, speech portals, travel reservation, stock quotes. The number of applications being written specifically for the car and for wireless devices is also increasing. Introducing a robust spoken dialogue interface to wireless terminals will enhance existing applications and help create new ones. High speech recognition accuracy for a variety of channel and noise conditions is essential for the success of ASR applications and services. Our goal in this paper is to investigate the degradation in speech recognition performance under typical wireless channel conditions and propose error protection and concealment strategies that improve performance. In addition to channel errors, adverse speech recording conditions can also degrade ASR performance, e.g., hands-free operation of end-devices or ambient noise [6], and are not treated in this paper.

In a distributed speech recognition system, a small client program running in the device extracts and transmits representative parameters of the speech signal from the mobile terminal over the wireless network to a multiuser speech recognition server. The alternative approach of performing speech recognition locally on the device significantly increases computation, power, and memory requirements for the device, and limits portability across languages and application domains. In this paper, we adopt the distributed approach to wireless ASR. Parameters that are optimized for speech recognition are extracted at the terminal, quantized at a source bit rate of 6 kb/s, and transmitted over the channel.

In [2], severe ASR performance degradation was observed for a distributed wireless speech recognition system, especially in the case of bursty transmission errors. In [7], specialized channel coding schemes for the transmission of ASR parameters were proposed. Specifically, a robust unequal error protection scheme with both forward error correction and detection capabilities to give a total bit rate of 9.6 kb/s was proposed. The error protection scheme was shown to significantly improve ASR performance over a wide variety of noisy wireless channels without the additional delay and bandwidth needed to retransmit the speech parameters.

In this paper, "soft-outputs" from the channel decoder are used to improve the performance of the speech recognition system. Specifically, the confidence level for each decoded bit is obtained using the Max-Log MAP algorithm and is used to estimate the confidence in ASR features and weight the contribution of each feature in the likelihood computation formula applied during decoding. This novel "soft-feature" decision is shown to significantly improve ASR performance.

2. SPEECH PARAMETERS AND QUANTIZATION

The acoustic features for speech recognition used in this study are the signal energy, e, and the 12 cepstral coefficients, $c_1, c_2, ..., c_{12}$, calculated every 10 ms based on a LPC analysis of order 10. The signal sampling rate is 8000 Hz and a Hamming window with 240 samples is used. These features form a 13-dimensional vector every 10 ms, which is the acoustic input to the automatic speech recognition system.

In order to transmit from the wireless handset to the network based recognition server, all 13 features are scalar-quantized. A simple non-uniform quantizer is used to determine the quantization cells. The quantizer uses the empirical distribution function as the companding function, so that samples are uniformly distributed in the quantization cells. Empirical tests showed no noticeable performance degradation when c_{12} is not transmitted. Therefore, we transmit only 12 feature components: energy, e, and $c_1 - c_{11}$. The bit allocation scheme for these feature components is shown in Table 1. The total number of bits for this bit allocation scheme is 60 bits per 10 ms frame. This

Feature Component	e, c_1, c_2, c_3 c_4, c_5	c_6, c_7, c_8 c_9, c_{10}, c_{11}	c_{12}
Bits per Feature	6	4	0

Table 1: Bits Allocation for Different Feature Components

requires an uncoded data rate of 6 kb/s to be transmitted over the wireless channel which will be the data rate used throughout this paper.

3. TRANSMISSION SYSTEM

In [7], several unequal error protection (UEP) schemes were proposed for the transmission of the quantized ASR features over noisy wireless channels. The performance of the speech recognizer under the proposed UEPs was evaluated for various channel types and conditions. In this paper, we concentrate on a single UEP scheme described below. Features from each 10ms speech frame are quantized into 60 source bits; with the addition of error protection bits the UEP coded data rate becomes 9.6 kbits/s. Binary phase shift keying (BPSK) is used. To provide better time diversity and improve performance in slow fading channels coded data is interleaved over 8 speech frames or 80 ms. The interleaving and deinterleaving delay associated with this is 160 ms and is tolerable for our application. The total number of coded bits in an 80 ms channel encoded frame is 768.

The UEP scheme consists of 3 levels denoted by L1, L2, and L3; furthermore, L1 is separated to two levels L1_1 and L1_2. The assignment of the bits for different UEP levels is shown in Table 2. In this notation, e^0 denotes the MSB of e. As seen from the table the number of bits per speech frame in L1, L2, and L3 are 13, 24 and 23, respectively. L1_1 contains the bits that are determined to be the most important 7 bits. L1_2 contains the next 6 important bits. A rate 1/2 memory 5 convolutional code is used on L1 and L2 bits. Channel coding is done so that L1_1 bits are followed by L1_2 and then L2. Note that because of the punctured code used with L2 bits those bits of L1_2 that are within a decoding depth of L2 bits will not be subjected to the usual rate 1/2 mother code. The L1_2 bits, in a channel coded frame of 80 ms, are arranged in the following manner: $e^2(n), e^2(n+1), ..., e^2(n+7)$; $c_1^1(n), c_1^1(n+1), ..., c_1^1(n+7)$; $c_5^1(n), c_5^1(n+1), ..., c_5^1(n+7)$. We have determined experimentally that the coefficients $c_1(n)$ are more significant than $c_5(n)$ and, therefore, this bit arrangement will assign a gradually decreasing error protection level to those coefficients that are toward the end of the L1_2 frame. The total coded bits in 8 speech frames from L1 bits is 208. For the 197 L2 bits (including the 5-bit tail) we use a rate 1/2 code with 18 bits punctured to give 376 coded bits. Then, with the 184 L3 uncoded bits the total coded bits in 8 speech frames is 768.

Level	Speech Bits	Error Protection
L1_1	$e^0, e^1, c_1^0, c_2^0, c_3^0, c_4^0, c_5^0$	rate 1/2 conv. code
L1_2	$e^2, c_1^1, c_2^1, c_3^1, c_4^1, c_5^1$	rate 1/2 conv. code
L2	$e^3, e^4, c_1^2, c_1^3, ..., c_5^2, c_5^3$ $c_6^0, c_6^1, c_7^0, c_7^1, ..., c_{11}^0, c_{11}^1$	rate 1/2 conv. code and puncturing
L3	$e^5, c_1^4, c_1^5, ..., c_5^4, c_5^5$ $c_6^2, c_6^3, c_7^2 c_7^3, ..., c_{11}^2, c_{11}^3$	no code

Table 2: Speech bit assignment for different UEP levels in UEP1.

3.1. Bit Error Probabilities

At the receiver we employ coherent demodulation with perfect channel state information and use an algorithm that gives the *a posteriori probability* (APP) for each decoded bit. The Max-Log-MAP algorithm [5] gives the approximate APPs or the log likelihood $\Lambda(n) = \ln \frac{prob(\hat{a}(n)=1)}{prob(\hat{a}(n)=0)}$, where $\hat{a}(n)$ is the channel decoder output. Note that larger values of $|\Lambda(n)|$ increases the reliability measure of the channel decoder output, $\hat{a}(n)$. In this work, we employ a one-bit quantization of the likelihood ($\hat{\Lambda}(n)$) to a reliability measure of $\hat{a}(n)$ and the source decoder receives the decoded data symbol together with its one-bit reliability measure. This reliability is derived as follows. Denote by $\Lambda_T (> 0)$ a predetermined threshold, then, if $|\Lambda(n)| < \Lambda_T$ $\hat{\Lambda}(n) = 1$; else $\hat{\Lambda}(n) = 0$. That is, $\hat{\Lambda}(n) = 1$ signifies a potential error and therefore signals the source decoder to consider the relevant decoded bit as an erasure. In Figure 1, the 'false accept' and 'false reject' rates (a) and absolute number (b) are shown for various thresholds. 'False accept' signifies accepting as correct a bit that was decoded erroneously and 'false reject' denotes erroneously rejecting a correctly decoded bit. Note that the *equal error rate* is achieved around $\Lambda_T = 3$, while the *equal number of errors* point is around $\Lambda_T = 1$. Similar results were obtained for channel conditions different than the ones used for Fig. 1.

4. SOFT-FEATURE DECODING

To overcome the detrimental effects of transmission errors, common error concealment strategies include the repetition of previously received frames or parameter interpolation. These techniques may help to repair random bit errors but may fail for errors occurring in bursts, which are very likely in fading channels. In this section, we consider a novel error concealment technique which is based on "soft-outputs" from the channel decoder The *a posteriori* probability of each decoded bit is produced at the channel decoder as discussed in the previous section, and is then utilized by the ASR decoder to improve performance.

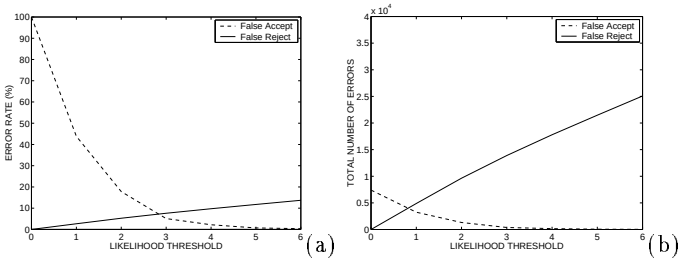

Figure 1: The percent (a) and absolute number (b) of type I and type II errors for various likelihood ratio thresholds Λ_T. A 3 dB SNR Rayleigh fading channel at 10 km/h was used for this simulation.

4.1. Marginalize Unreliable Features

The first proposed error concealment strategy discards the transmitted features which are most probably erroneous and uses only the reliable ones for likelihood computations at the speech recognizer. A reduced feature vector is used based only on the components that have a high confidence level. In an hidden Markov model (HMM) based speech recognition system, the observed feature vectors are modeled by state-specific probability distributions $p(x|s)$, where x is the feature vector and is the state of the model. Usually a mixture of Gaussian densities is used for each state of the phoneme (or triphone) specific HMM. In this case, the reduced distribution for the reliable part of the feature vector is the marginal determined by integrating over all unreliable components:

$$p(x_{rel}|s) = \int p(x|s)dx_{unrel}. \quad (1)$$

where x_{rel}, x_{unrel} are the reliable and unreliable components of the feature vector. Using the distribution of only the reliable components for HMM likelihood computation is one of the techniques for improving robustness of speech recognizers in noisy conditions, often labeled as the "missing feature theory" [1]. For speech recognition in noise, labeling unreliable spectral features can be a challenging task, while in our application the reliability of each feature is provided by the channel decoder. With diagonal covariance Gaussian mixture modeling, the reduced likelihood function can be easily calculated by dropping unreliable components from the full likelihood computation [1]. This approach requires little modification in existing speech recognition systems. Alternatively, feature components in the likelihood computation can also be weighted by their confidence values. In this case, continuous confidence values between 0 and 1 would be used and the contribution of each feature to the likelihood computation would be scaled by its confidence as discussed in Section 4.2.

The soft-feature decoding algorithm that computes likelihood using the marginal distribution over unreliable features will be henceforth refered to as **SoftFeatI** and is implemented as follows: (i) for energy and cepstrum features, if the first or the second bit of the decoded symbol has absolute likelihood ratio $|\Lambda(n)|$ below the threshold Λ_T it is labeled as 'unreliable' and not used in the likelihood computation (marginalize according to Eq. (1)), (ii) for 'delta' and 'delta-delta' features (smooth first and second derivatives of the energy and cepstrum features), if the first or the second bit of any of the symbols in the window used for the delta computation has $|\Lambda(n)| < \Lambda_T$, then, do not use the delta feature in the likelihood computation. Five and seven frame windows are used for the delta and delta-delta computation, respectively. The likelihood ratio threshold Λ_T that minimizes recognition error can be computed from held out data.

4.2. Exponential Feature Weights

An alternative soft-feature decoding algorithm, labeled **SoftFeatII** applies exponential weights to each feature in the probability computation at the decoder. Specifically, assuming that the state observation probability density function (pdf) is a mixture of Gaussian pdfs with diagonal covariance the observation probability computation formula is modified as follows:

$$p(x|s) = \sum_{m=1}^{M} w_m \prod_{n=1}^{N} \left[\frac{1}{\sqrt{2\pi}\sigma_{nm}} \exp\left(-\frac{(x_n-\mu_{nm})^2}{2\sigma_{nm}^2}\right)\right]^{f(C_n)} \quad (2)$$

where x is the feature vector, N is the size of the feature vector, M is the number of Gaussian mixtures per state and w_m, μ_m, σ_m are the mixture weight, mean and standard deviation, respectively, of the mth Gaussian for HMM state s. C_n is the confidence associated with the nth feature and $f(C_n)$ is a function of the confidence C_n. Note that C *is a function of time* and is updated at the frame rate, i.e., as often as x is updated. Assuming that the confidence is normalized to a number between 0 and 1, then one possible form of the function $f(C)$ is $f(C) = (\alpha + C)/(\alpha + 1)$ where α is a smoothing constant that is experimentally determined so that error is minimized on a held-out data set. For very large values of α, all features are more or less weighted equally (confidence C is practically ignored), while for very small values of α, only features with high confidence ($C_n \approx 1$) are considered in the observation probability computation. All other aspects of the decoding process, apart from the feature weighting in the state observation probability computation, remain unchanged.

To compute the feature confidence the symbol bit probabilities computed at the channel decoder are translated into feature confidence scores. In our case, the mapping from feature value to sequence of bits (quantization) is nonlinear. The feature confidence score is thus computed numerically as follows. Assuming that the most probable symbol obtained at the channel decoder is $\hat{S}$, and $P(S_k)$ is the probability that the decoder produces the kth symbol[1], the expected mean square error E for that feature is computed as:

$$E = \sum_k P(S_k)\left[Q^{-1}(\hat{S}) - Q^{-1}(S_k)\right]^2 \quad (3)$$

where Q^{-1} is the inverse of the quantization mapping. The expected mean square error is normalized by the feature variance and subtracted from 1 to produce the feature confidence C.

5. EXPERIMENTAL RESULTS

The performance of the ASR system for various transmission channel conditions and error protection schemes was evaluated on an isolated word speech recognition task, where people were asked to spontaneously answer questions about their mother language, country of birth etc. The database was collected over the public telephone network. A total of 4387 utterances were used for the system evaluation. The vocabulary size was 23 different words. This test set consists of speakers from all over the United States with large dialect diversity and a significant number of non-native speakers.

The 12 LPC-derived cepstral coefficients, the signal energy, and their 1st and 2nd order time derivatives were used as acoustic features for speech recognition. The acoustic models for speech recognition were trained on a collection of speech databases collected over the public telephone network. The speech recognizer is based on continuous density

[1]Assuming independence among bits, $P(S)$ is simply computed as the product of the decoder probabilities for each of the bits in the symbol.

HMMs and the Bell Labs recognition engine. The acoustic units are state-clustered triphone models, having three emitting states and a left-to-right topology.

5.1. Quantization

The baseline word error rate for this task (on the unquantized data) was 6.8 %. The relatively high error rate is due to the noisy conditions, the usage of speaker-phones, and hesitations and filled pauses in the data (spontaneous speech). Using the proposed 60 bits per speech frame bit allocation scheme the error rate increases slightly to 7.2 %. More complex quantization schemes can achieve loss-less (for speech recognition purposes) compression of the ASR feature set at bit rates lower than 6 kb/s. For example in [4], a loss-less vector quantization scheme is proposed that operates at 4 kb/s. Note that soft-feature decoding can also be used with a vector quantization scheme.

5.2. Speech Recognition Results

In the first set of experiments, we study the choice of the likelihood ratio threshold Λ_T for the SoftFeatI algorithm. In Table 3, word error rate is shown as a function of the threshold. It is clear from the table that SoftFeatI gives optimum performance for Λ_T around 1, which is the same region where equal number of false accept and false reject errors is achieved in Fig 1(b). Note that when the threshold is 0 all features are assumed to be correctly transmitted and used in the likelihood computation (this is the baseline performance). Also as the threshold reaches the equal error rate region (around 3 in Fig 1(a)) results become worse than the baseline. Overall, a choice of the threshold that achieves an equal *number* of accept and reject errors produces the best recognition results for the SoftFeatI algorithm.

In the next set of experiments, we evaluate soft-feature decoding performance for a *simulated* correlated fading channel with the channel correlation given by the mobile speed. Word error rates for channels with various mobile speeds and SNRs are listed in Table 4. Results are shown with and without soft-feature decoding. Both SoftFeatI and SoftFeatII algorithms are evaluated. A likelihood ratio threshold of 1 is used for the SoftFeatI experiments. Significant improvements are shown over baseline for both SoftFeatI and SoftFeatII algorithms. Specifically for the SoftFeatI algorithm, relative error rate reduction of 20-30% is shown for noisy channel conditions. In general, higher improvement is shown for faster speed and lower SNR channels. The improvements are impressive and are achieved at nominal additional computational cost. The SoftFeatII algorithm further improves recognition performance and reduces error rate by 30-50% over the baseline. The improvement is substantial and is equivalent to enhancing the channel SNR by about 1.5 dB. Similar results have been obtained for a Gaussian wireless channel, e.g., at -2 dB SNR for a Gaussian channel ASR error rate reduces from 26.6% to 12.5% using the SoftFeatII algorithm. More research is underway to tune the parameters of the SoftFeatII algorithm on held-out data to further improve performance.

	Likelihood Ratio Threshold				
	0	1	2	3	5
word error rate (%)	32.5	23.9	29.9	36.3	47.7

Table 3: SoftFeatI word error rates (%) for various thresholds (0 dB SNR Rayleigh fading channel at 50 km/h). Note that $\Lambda_T = 0$ is the baseline ASR performance.

Speed [km/h]	Decoding Scheme	SNR 5 dB	3 dB	1.5 dB	0 dB
10	baseline	11.2	17.2	24.8	37.7
	SoftFeatI	9.9	14.6	19.4	30.3
	SoftFeatII	9.2	11.8	16.2	25.6
50	baseline	8.1	11.4	17.6	32.5
	SoftFeatI	7.8	9.7	13.1	23.9
	SoftFeatII	7.6	8.5	10.5	16.7
100	baseline	7.5	9.3	14.1	30.7
	SoftFeatI	7.4	8.5	12.0	21.4
	SoftFeatII	7.7	8.0	9.1	14.5

Table 4: Word error rates with and without soft-feature decoding for a Rayleigh fading channel with different speeds and SNRs. "SoftFeatI" denotes marginalizing over unreliable feature (likelihood ratio threshold is 1) while "SoftFeatII" denotes weighting by feature confidence ($a = 0$). Bit error probabilities are estimated at the channel decoder using the Max-Log-MAP algorithm.

6. CONCLUSIONS

A novel soft-feature decoding algorithm was proposed for speech recognition over wireless channels. The algorithm employs the bit probabilities at the channel decoder to assign confidence on the ASR features. This information is used during decoding to improve recognition performance. Up to 50% relative error rate reduction is shown for certain channel conditions. Improvements are shown under all channel conditions with minimal additional computational load. The proposed algorithms can enhance speech recognition performance for any wireless channel, ASR feature encoding scheme and channel protection scheme.

7. REFERENCES

[1] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust ASR with Unreliable Data and Minimal Assumption," in *Proceedings, Robust Methods for Speech Recognition in Adverse Conditions*, (Tampere, Finland), pp. 195–198, 1999.

[2] A. Gallardo-Antolin, F. D. de Maria, and F. Valverde-Albacete, "Avoiding Distortions Due to Speech Coding and Transmission Errors in GSM ASR Tasks," in *1999 International Conference on Acoustics, Speech and Signal Processing*, (Phoenix, Arizona), 1999.

[3] P. Haavisto, "Speech Recognition for Mobile Communications," in *Proceedings, Robust Methods for Speech Recognition in Adverse Conditions*, (Tampere, Finland), pp. 15–18, 1999.

[4] G. N. Ramaswamy and P. S. Gopalakrishnan, "Compression of Acoustic Features for Speech Recognition in Network Environments," in *1998 International Conference on Acoustics, Speech and Signal Processing*, (Seattle, Washington), 1998.

[5] P. Robertson, P. Hoeher, and E. Villebrun, "Optimal and Sub-Optimal Maximum *A Posteriori* Algorithms Suitable for Turbo-Decoding," *European Trans. Telecommun. (ETT)*, vol. 8, pp. 119–125, March/April 1997.

[6] F. Soong and E. Woodenberg, "Hands-Free Human-Machine Dialogue, Corpora, Technology and Evaluation, in *2000 International Conference on Speech and Language Processing*, (Beijing, China), 2000.

[7] V. Weerackody, W. Reichl, and A. Potamianos, "An Error-Protected Speech Recognition System for Wireless Communications," submitted to *IEEE JSAC: Wireless Communications*, 2000.