CHANNEL NOISE ROBUSTNESS FOR LOW-BITRATE REMOTE SPEECH RECOGNITION

Alexis Bernard and Abeer Alwan

Department of Electrical Engineering, University of California, Los Angeles
{abernard, alwan}@icsl.ucla.edu

ABSTRACT

In remote (or distributed) speech recognition , the recognition features are quantized at the client, and transmitted to the server via wireless or packet-based communication for recognition. In this paper, we investigate the issue of robustness of remote speech recognition applications against channel noise. The techniques presented include: 1) optimal soft decision channel decoding allowing for error detection, 2) weighted Viterbi recognition (WVR) with weighting coefficients based on the channel decoding reliability, 3) frame erasure concealment, and 4) WVR with weighting coefficients based on the quality of the erasure concealment operation. The techniques presented are implemented at the receiver (server), which limit the complexity for the client, and significantly extend the range of channel conditions for which remote recognition can be sustained. As a case study, we illustrate that remote recognition based on perceptual linear prediction (PLP) coefficients is able to provide at less than 500 bps, good recognition accuracy over a wide range of channel conditions.

1. INTRODUCTION

Wireless communications is a challenging environment for remote speech recognition. The communication link is characterized by time-varying, and sometimes low signal-to-noise ratio (SNR), channels. Previous studies have suggested alleviating the effect of channel errors by adapting acoustic models [1] or automatic speech recognition (ASR) front-ends [2] to different channel conditions .

Similarly, packet switched communication networks also constitute a difficult environment for remote recognition applications. The communication link in IP based systems is characterized by packet losses, mainly due to congestion at routers. Packet loss recovery techniques including silence substitution, noise substitution, repetition and interpolation were presented in [3].

We investigated the effect of channel erasures and errors on remote recognition accuracy in [4] for isolated digit recognition based on PLP coefficients and in [5] for continuous digit recognition based on MFCCs. Both papers showed that channel errors, which propagate through the trellis, have a disastrous effect on recognition accuracy, even at less than 1%, while the recognizer is able to operate with almost no loss in accuracy with up to 10% of channel erasures.

In [4], we presented a sub-optimal technique for combining the advantages of soft decision decoding and error detection by computing the log ratio of the *a posteriori* probabilities of the two most likely codevectors. We present in this paper the optimal solution for soft decision based error detection, which is based on the log likelihood of the two most likely codevectors.

In [5, 6], the Viterbi recognizer is modified to include a time-varying weighting factor depending on the reliability of each decoded feature. In this paper, we illustrate how the weighting coefficients can be derived from the channel decoder or the frame erasure concealment operation. For the first case, a technique for computing the decoding reliability based on the soft received bits is presented. For the second case, the weighting coefficients are derived from the quality of the concealment operation.

Section 2 presents an optimal technique for combining soft decision decoding with error detection. Section 3 presents the weighted Viterbi recognition (WVR) algorithm that takes into account the reliability of the features. Section 4 presents several techniques aimed at alleviating the effect of channel erasures. Section 5 compares recognition results using the different techniques.

2. OPTIMAL SOFT DECISION BASED ERROR DETECTION

When designing channel coders and decoders for remote recognition applications, the emphasis should be on error detection more so than on error correction [4].

Soft decision decoding always outperforms hard decision decoding, both for AWGN and multi-path communication channels. However, classic soft decision decoding does not provide a tool for error detection and is not suited for remote speech recognition applications governed by the criteria of low probability of undetected error. We present here a method for combining soft decision decoding with error detection.

Since maximum likelihood is the optimal decision rule for decoding channels with Gaussian statistics (noise variance $\sigma^2 = N_0/2$), it is desirable to perform error detection based on the ratio of the likelihoods of the two most probable codevectors. Using Bayes rule and assuming that all codewords are equiprobable, the ratio of the log likelihoods of the two most probable vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ (which are also the two closest codevectors from the received vector $\boldsymbol{y}$ at Euclidean distances d_{E_1} and d_{E_2} from $\boldsymbol{y}$) is given by

$$\frac{P(\boldsymbol{y}|\boldsymbol{x}=\boldsymbol{x}_1)}{P(\boldsymbol{y}|\boldsymbol{x}=\boldsymbol{x}_2)} = e^{\frac{1}{2}\frac{(d_{E_2}^2 - d_{E_1}^2)}{\sigma^2}}. \quad (1)$$

If one determines the projection of the received codevector $\boldsymbol{y}$ onto the line segment joining $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ as defining the distance d_1 and d_2 on the inter-segment (Fig. 1), triangle geometry tells us that $(d_{E_2}^2 - d_{E_1}^2) = (d_2 - d_1)(d_2 + d_1)$ and Eq. 2 can be rewritten as

$$\frac{P(\boldsymbol{y}|\boldsymbol{x}=\boldsymbol{x}_1)}{P(\boldsymbol{y}|\boldsymbol{x}=\boldsymbol{x}_2)} = e^{\frac{1}{2}(\frac{D}{\sigma})^2(\frac{d_2-d_1}{D})} \quad (2)$$

This work was supported in part by HRL, ST Micro electronics and Broadcom, through the UC Micro program and by the NSF.

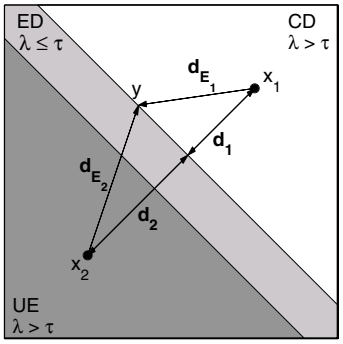

Fig. 1. λ-soft decision decoding for a (2,1) linear block code, assuming $\boldsymbol{x_1}$ was transmitted. CD, ED and UE stand for correct decoding, error detection and undetected error, respectively.

where $D = d_2 + d_1$ is the Euclidean distance between the two codevectors closest to the received codeword $\boldsymbol{y}$. The important factor in Eq. 2 is

$$\lambda = \frac{d_2 - d_1}{D}\,, \tag{3}$$

which is independent of the channel noise N_0 and represent a good indication of the channel decoding reliability.

If $\lambda = 0$, both codevectors are equally probable and the decision of the decoder should be rejected. If $\lambda = 1$, correct decision is almost guaranteed since the channel codes used are chosen according to channel conditions so that the minimum Euclidean distance between any two codevectors is at least several times as large as the expected noise ($D^2/N_0 \gg 1$). The soft decision-based error detection channel decoding scheme will be referred to as **$\boldsymbol{\lambda}$-soft** decoding. Fig. 1 illustrates the λ-soft decoding operation.

With λ-soft decoding, error detection can be declared when λ is smaller than a given threshold ($\lambda < \tau$). The probability of erasure declaration increases with λ. Classic soft decision decoding is a particular case with $\lambda = 0$.

3. WEIGHTED VITERBI RECOGNITION (WVR)

With remote recognition, reliability of the decoded features is a function of channel characteristics. When channel characteristics degrade, one can no longer guarantee the confidence in the decoded feature. The weighted Viterbi recognizer (WVR) in [5] presents a solution for modifying the recursive step of the Viterbi algorithm (VA) to take into account the effect of channel transmission by weighting the probability of observing the decoded feature given the HMM state model $b_j(\boldsymbol{o}_t)$ with the probability of decoding the feature vector $\boldsymbol{o}_t$. The time-varying weighting coefficient γ_t can be inserted into the VA by raising the probability $b_j(\boldsymbol{o}_t)$ to the power γ_t to obtain the following state update equation

$$\phi_{j,t} = \max_i \left[\phi_{i,t-1}\; a_{ij}\right] \left[b_j(\boldsymbol{o}_t)\right]^{\gamma_t}, \tag{4}$$

where $\phi_{j,t}$ is the state metric for state j at time t and a_{ij} is the state transition metric. Under the hypothesis of a diagonal covariance matrix $\boldsymbol{\Sigma}$, the overall probability $b_j(\boldsymbol{o}_t)$ can be computed as the product of the probabilities of observing each individual feature. The weighted recursive formula (Eq. 4) can include individual weighting factors $\gamma_{k,t}$ for each of the N_F front-end features:

$$\phi_{j,t} = \max_i \left[\phi_{i,t-1}\; a_{ij}\right] \prod_{k=1}^{N_F} \left[b_j(o_{k,t})\right]^{\gamma_{k,t}}. \tag{5}$$

4. ROBUSTNESS AGAINST CHANNEL TRANSMISSION

This section presents different techniques specifically designed for coping with channel transmission and erasures .

4.1. Frame dropping

The first technique reduces channel effects on recognition accuracy by detecting channel errors, and consequently removing the "suspicious" feature vectors from the sequence of observations [4]. The motivation behind this technique is that channel errors rapidly degrade recognition accuracy, while recognizers can cope with missing segments in the sequence of observations given the redundancy of the speech signal. The drawback of removing the unreliable frames from the stream of feature vectors is that the timing information associated with them is lost. This can significantly impact recognition accuracy.

4.2. WVR based on channel decoding reliability (λ-WVR)

We introduced the WVR technique in [5] to match the recognizer with the confidence in the decoded feature after channel transmission. We present here a new channel decoding reliability measurement based on the new λ-soft decision decoding scheme presented in Section 2. We consider both binary and continuous weighting.

With *binary* weighting, the weighting coefficients γ_t can either be zero (if the frame is lost or declared in erasure) or one (if the frame is received). The advantage of this technique over frame dropping, where state metrics are not updated ($\phi_{j,t} = \phi_{j,t-1}$), is that the timing information of the observation sequence is conserved. State metrics are continuously updated, even when $\gamma_t = 0$, by virtue of the state transition probability matrix using $\phi_{j,t} = \max_i \left[\phi_{i,t-1}\; a_{ij}\right]$.

The system can be refined if a time-varying *continuous* estimate $\gamma_t \in [0, 1]$ of the feature vector reliability is made available to the recognition engine. We propose the following square function to map the interval $[0, 1]$ for λ_t to the interval $[0, 1]$ for γ_t: $\gamma_t = \lambda_t^2$. The quadratic exponent is empirically chosen after it was shown to provide the necessary statistical rejection of the uncertain frames.

Note that if hard decision decoding was used, only binary weighting could be used. For soft decision decoding, one can choose to apply binary weighting with $\gamma_t = 0$ if $\lambda_t < \tau$ and $\gamma_t = 1$ if $\lambda_t \geq \tau$, or continuous weighting with $\gamma_t = \lambda_t^2$. Note that $\tau = 0.16$ was shown to be a sensible choice for τ.

4.3. Repetition-based frame erasure concealment

The problem when a large number of frames is not received at the decoder is that the synchronization of the Viterbi recognizer may be disturbed, even if state metrics are continuously updated using only the transition matrix. Hence, subsequent received features might be analyzed using an inappropriate state. This problem becomes more significant when erasures occur in bursts, almost forcing the best path in the trellis to remain in the same state for a long period of time.

GILBERT CHANNEL STATE	GOOD	BAD
Static features	$\gamma_{k,t} = \sqrt{\rho_k(t-tc)}$	
Dynamic features	$\gamma_{k,t} = 1$	$\gamma_{k,t} = 0$

Table 1. Determination of the frame erasure concealment based weighting coefficients for WVR.

Feature concealment methods not only preserve the timing information, but also attempt to recreate the missing feature vector by replacing it with an estimate. *Repetition*-based schemes replace missing frames with copies of previously-received frames [3].

In our scheme, the *static* features (front-end temporal features) of a missing frame are replaced with a copy of the front-end features of the last correctly received frame.

However, note that erasures propagate through the computation of the *dynamic* features (derivatives and accelerations). For this reason, only immediate left and right neighboring frames are used in our scheme for the computation of dynamic features. While this might result in a slight loss of performance in erasure-free conditions, it provides robustness against erasures.

The dynamic features of a missing frame are then computed at the receiver as follows. The receiver determines what the status of the channel is. If two consecutive frames are lost/received, the receiver determines that the channel is bad/good. If the channel is bad, the dynamic features are not computed and will be discarded in the Viterbi search by assigning zero weighting. If the channel is good, dynamic features are computed, with, if necessary, the use of one-sided derivatives if one of the two neighboring frame is missing.

4.4. WVR based on erasure concealment quality (ρ-WVR)

Performance of repetition concealment degrades rapidly as the number of consecutive lost frames increases. For instance, when packet losses approach or exceed the length of a phoneme (10-100 ms or 1-10 frames), the speech signal may already have evolved to another sound, which no longer justifies repetition of the last correctly received feature vector.

This section presents an extension to the repetition-based concealment technique, whereby the confidence in the frame erasure concealment is fed into the Viterbi recognizer for improved recognition performance. Indeed, it is beneficial to decrease the weighting factor $\gamma_{k,t}$ when the number of consecutively repeated frames increases. For the weighting coefficients of the static front-end features, we propose

$$\gamma_{k,t} = \sqrt{\rho_k(t-t_c)}, \tag{6}$$

where ρ_k is the time auto-correlation of the k^{th} feature and t_c is the time instant of the last correctly received frame. Note that if there is no erasure, $t = t_c$ and $\gamma_{k,t} = 1$. For the dynamic features, we use $\gamma_t = 1$ if we are in a good state and $\gamma_t = 0$ if we are in a bad state. Table 1 summarizes the WVR weighting coefficients as a function of the channel status.

Finally, note that the option of concealing the derivative features is not chosen given their time-correlation significantly less than that of the front-end features. Hence, repetition of dynamic features does not necessarily lead to a good estimate of the missing features and should be avoided.

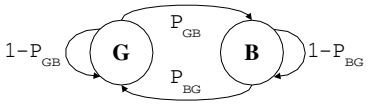

Fig. 2. State diagram for the Gilbert-Elliot bursty channel.

(P_{GB}, P_{BG})	(2.5,20)	(2.5,15)	(5,20)	(2.5,10)	(1.25,5)	(5,15)	(10,20)	(5,10)
P_B	11.1	14.3	20.0	20.0	20.0	25.0	33.3	33.3
P_E	17.8	20.0	24.3	24.3	24.3	27.5	33.3	33.3
$\overline{L_b}$	5.0	6.6	5.0	10.0	20.0	6.6	5.0	10.0

Table 2. Characteristics of the Gilbert-Elliot channels of interest. Probabilities are given in percent.

5. RECOGNITION RESULTS

This section compares recognition results for the different techniques aimed at alleviating the effect of channel erasures. The recognition experiment is based on continuous digit recognition using PLP features and the Aurora-2 database and its HMM based configuration (word models, 16 states and 6 mixtures/states).

Two types of erasure models are analyzed. In the first type, channel erasures occur independently, with a given probability of erasures. In the second type, channel erasures occur in bursts. This is typically the case for wireless or IP based communication systems, where correlated fading or network congestion may cause the loss of consecutive frames.

A classic model for bursty channels is the Gilbert-Elliot [7] model in which the transmission is modeled to be a Markov system where the channel is assigned to be in either one of two states: "0" for *good* and "1" for *bad*. Figure 2 illustrates a Gilbert channel model. With such a model, there is a probability $P_G = \frac{P_{BG}}{P_{BG}+P_{GB}}$ to be in state "0" and a probability $P_B = \frac{P_{GB}}{P_{GB}+P_{BG}}$ to be in state "1". If the probabilities of channel erasures are assigned to be P_{E_G} for the good state and P_{E_B} for the bad state, the overall average probability of erasure is: $P_E = P_G P_{E_G} + P_B P_{E_B}$.

Throughout the experiments, P_{E_G} is considered to be equal to 10% and P_{E_B} is set to 80%. Different types of bursty channels are analyzed, depending on the state transition probabilities P_{GB} and P_{BG}, which in turn determine how bursty the channel is. Table 2 summarizes the properties of the bursty channels studied, including the probability (in percent) of being in the bad state (P_B), the overall probability of erasure (P_E), and the average length (in frames) of a burst of erasures ($\overline{L}_b$).

5.1. Comparison of the different channel robust techniques

Table 3(a) illustrates recognition accuracy for the different frame erasure concealment techniques applied to the independent erasure channel. Baseline recognition accuracy for erasure-free channels is 98.52%. Some observations can be made. 1) After about 10-20% of independent frame erasures, recognition accuracy degrades rapidly. 2) Transmission of the binary frame erasure reliability measurement to the weighted Viterbi recognizer preserves synchronization of the VA and significantly reduces the word error rate. 3) Repetition-based frame erasure concealment, which in addition to preserving the timing also provides an approximation for the missing frame, typically outperforms binary λ-WVR. 4) Addition of the weighting coefficients $\gamma_{k,t}$ representing the quality of the feature concealment

Independent Erasures	0%	10%	20%	30%	40%	50%	60%
Frame dropping	98.52	97.19	93.51	85.49	71.23	56.33	38.76
Binary λ-WVR	98.52	98.31	98.11	97.19	96.87	94.31	93.19
Concealment	98.52	98.47	98.31	98.19	97.67	96.35	94.31
Concealment + ρ-WVR	98.52	98.52	98.47	98.39	98.11	97.61	96.01

(a) Independent erasure channels.

Gilbert Channels	(2.5,20)	(2.5,15)	(5,20)	(2.5,10)	(1.25,5)	(5,15)	(10,20)	(5,10)
Frame dropping	90.68	87.04	85.79	81.67	80.07	79.33	74.70	69.85
Binary λ-WVR	97.35	96.27	96.20	94.53	93.69	95.06	94.85	92.82
Concealment	97.41	96.41	96.77	94.42	93.27	94.35	93.83	92.11
Conc. + ρ-WVR	98.07	97.55	97.84	97.37	97.03	97.15	96.87	96.09

(b) Bursty (Gilbert) erasure channels.

Table 3. Recognition accuracy with the Aurora-2 database and PLP features with derivatives using two types of channel erasures: (a) independent and (b) bursty. Different techniques are compared: frame dropping; frame dropping with binary λ-WVR ($\gamma_t = 0$ if frame is dropped); frame erasure concealment (repetition); and concealment with ρ-WVR ($\gamma_{k,t} = \sqrt{\rho_k(t-t_c)}$).

technique (Eq. 6) in the Viterbi search further improves recognition performance.

These results are confirmed in Table 3(b) for the bursty Gilbert channels of Table 2 for which we can make additional observations. 1) Binary WVR may outperform repetition-based erasure concealment when the average burst lengths are large. 2) Again, frame erasure concealment combined with WVR provides the best recognition results. For instance, for the Gilbert channel with $(P_{GB}, P_{BG}) = (1.25, 5)$, recognition accuracy improves from 93.27% to 97.03%, a 71% relative word error rate (WER) reduction compared to the baseline recognition performance of 98.52%. 3) Despite average overall probability of frame erasures between 18% and 33% and average length of erasure bursts between 5 and 20 frames (see Table 2), recognition accuracy is kept within 2% of the baseline erasure-free performance.

Note that Table 3 does not include results for continuous λ-WVR, which require simulations of a complete remote recognition system, including channel coding and decoding. Given the superior performance of ρ-WVR in Table 3 over the other techniques, we compare in Section 5.2 the performance of continuous ρ-WVR and λ-WVR on a complete remote recognition system.

5.2. Comparison between continuous λ-WVR and ρ-WVR

Table 4 presents recognition accuracy of a complete remote recognition system (source and channel coding, channel decoding) over a wide range of independent Rayleigh fading channels. Source coding is applied on the LSFs of the PLP system, with 5 to 7 bits per 20 ms frame, using the technique proposed in [4]. Depending on the channel conditions, different linear block codes maximizing error detection are used for channel protection [4]. The overall bit rate, including source and channel coding, is limited to 500 bps.

Two scenarios are considered. In the first scenario (λ-WVR), all the features are transmitted to the recognizer, even the unreliable ones, and the weighting coefficients ($\gamma_t = \lambda_t^2$) will lower the importance of the inaccurate ones. In the second scenario (ρ-WVR), the unreliable features (those for which $\lambda_t \leq 0.16$) are dropped and concealed with a substitution feature vector. The WVR weighting coefficient is based on the quality of the concealment operation

Block Code (N,K)	Bit Rate (bps)	SNR (dB)	BER (%)	RECOGNITION (%) Cont. λ-WVR $\gamma_t = \lambda_t^2$	ρ-WVR $\gamma_{k,t} = \sqrt{\rho_k}$
(8,7)	400	9	2.88	98.5	98.5
(8,6)	400	8	3.55	98.3	98.2
(8,6)	400	7	4.35	98.3	98.3
(10,7)	500	6	5.30	98.4	98.5
(10,7)	500	5	6.42	98.2	98.3
(10,6)	500	4	7.71	98.1	98.1
(10,6)	500	3	9.19	97.6	97.7
(10,5)	500	2	10.85	97.4	97.6

Table 4. Comparison between performance of channel based continuous ($\gamma_t = \lambda_t^2$) and erasure concealment based WVR ($\gamma_{k,t} = \sqrt{\rho_k(t-t_c)}$) over independent Rayleigh fading channels.

($\gamma_{k,t} = \sqrt{\rho_k(t-t_c)}$). Table 4 indicates that no one strategy always outperforms the other in a statistically significant manner for independent Rayleigh fading channels.

6. SUMMARY

We presented in this paper several techniques that provide robustness to remote speech recognition applications against channel noise. First, we present an optimal technique for combining soft decision decoding with error detection. Second, we present different techniques alleviating the effect of channel transmission: frame dropping, weighted Viterbi recognition based on the channel decoding reliability (λ-WVR), repetition-based frame erasure concealment, and WVR based on the quality of the concealment operation (ρ-WVR). Reductions in word error rates offered by the different techniques are analyzed. Note that the proposed solutions are implemented at the server, and does not increase the complexity to the client. It is shown that using such techniques combined with source and channel coding suitable for remote speech recognition, good recognition accuracy can be obtained over a wide range of channel conditions at bit rates less than 500 bps.

7. REFERENCES

[1] T. Salonidis and V. Digalakis, "Robust speech recognition for multiple topological scenarios of the GSM mobile phone system," in *Proc. of ICASSP*, May 1998, pp. 101–4.

[2] S. Dufour, C. Glorion, and P. Lockwood, "Evaluation of the root-normalised front-end for speech recognition in wireless GSM network environments," in *Proc. of ICASSP*, May 1996, vol. 1, pp. 77–80.

[3] B. Milner, "Robust speech recognition in burst-like packet loss," in *Proc. of ICASSP*, May 2001, vol. 1, pp. 261–4.

[4] A. Bernard and A. Alwan, "Source and channel coding for remote speech recognition over error-prone channels," in *Proc. of ICASSP*, May 2001, vol. 4, pp. 2613–6.

[5] A. Bernard and A. Alwan, "Joint channel decoding - Viterbi recognition for wireless applications," in *Proceedings of Eurospeech*, Sept. 2001, vol. 4, pp. 2703–6.

[6] A. Potamianos and V. Weerackody, "Soft-feature decoding for speech recognition over wireless channels," in *Proc. of ICASSP*, May 2001, vol. 1, pp. 269–72.

[7] E.N. Gilbert, "Capacity of burst noise channel," *Bell System Technical Journal*, vol. 39, pp. 1253–65, Sept. 1960.