Noise-Robust speech recognition of Conversational Telephone Speech

Gang Chen, Hesham Tolba and Douglas O'Shaughnessy
Université du Québec, Canada
{gangchen, tolba, dougo}@emt.inrs.ca

10.21437/Interspeech.2006-338

Abstract

Over the past several years, the primary focus of investigation for speech recognition has been over the telephone or IP network. Recently more and more IP telephony has been extensively used. This paper describes the performance of a speech recognizer on noisy speech transmitted over an H.323 IP telephony network, where the minimum mean-square error log spectra amplitude (MMSE-LSA) method [1,2] is used to reduce the mismatch between training and deployment condition in order to achieve robust speech recognition. In the H.323 network environment, the sources of distortion to the speech are packet loss and additive noise. In this work, we evaluate the impact of packet losses on speech recognition performance first, and then explore the effects of uncorrelated additive noise on the performance. To explore how additive acoustic noise affects the speech recognition performance, seven types of noise sources are selected for use in our experiments. Finally, the experimental results indicate that the MMSE-LSA enhancement method apparently increased robustness for some types of additive noise under certain packet loss rates over the H.323 telephone network.

Index Terms: speech recognition, H.323, telephone speech.

1. Introduction

Currently the rapidly increasing use of both the Internet telephony and Automatic Speech Recognition (ASR) makes Internet-telephony-based speech recognition extremely attractive. These kinds of schemes are based on a client-server recognition system, i.e., the client device generates quantization and packetization of speech signals and transmits them over the IP (internet protocol) telephony channel to a remote ASR server, which performs speech recognition.

Contrary to the switched telephone network, IP networks are not intended to transmit voice. This scenario yields some problems, which are packet loss, delay, and network jitter, to obstruct the deployment of VoIP. Usually, routers discard packets on a congested IP channel as soon as their packet in-flow surpasses their out-flow in a given route. Because full segments of the speech signal can be lost during transmitting speech data over IP, an important factor, packet loss, has to be taken into consideration when designing an ASR system.

The H.323 [3] is the leading standard in voice over IP (VoIP), and is a set of protocols for voice, video, and data conferencing over packet-based networks. Therefore, we built an H.323 channel environment to simulate a real H.323 telephone channel of the speech transmission for evaluating the influence of missing speech packets on robust ASR system performance.

Generally, a speech recognizer trained on clean speech data and working in adverse conditions has decreased performance, and clean speech utterances can be corrupted by additive noise. For mitigating speech distorted by additive noise, the MMSE-LSA method is used to enhance the speech corrupted by some kinds of additive noise. In this method, the short-time spectral amplitude (STSA) of the speech signal is estimated and combined with the short-time phase of the degraded noisy speech, to construct the robust signal.

2. Speech transmission

2.1. Implementation of voice over IP

Voice over Internet Protocol (VoIP [4]) is a means to talk over IP rather than solely over the Public Switched Telephone Network (PSTN). VoIP can be implemented by several methods. The first case is a voice call made from one PSTN telephone to another. This call can either be transmitted over traditional analog lines, or can be converted to IP, then back to the PSTN. The second instance depicts a voice call made on a PSTN telephone to a personal computer. The next scenario is a voice call initiated from a PC (personal computer) via its VoIP server, acting in a PSTN capacity, which is routed over the Internet to a telephone. Finally, the fourth scenario illustrates simply a PC-to-PC call where the voice signal is transported via IP without accessing the PSTN. To examine the overall impact of missing speech packets and additive noise over IP telephony on the ASR performance, we did not only simulate the PC-to-PC where the voice signal is transported via IP without accessing the PSTN by using the TIMIT database, but also simulated other situations, which are the PSTN telephone to another, the PSTN telephone to PC, and PC to PSTN telephone by transmitting speech data of the NTIMIT corpus.

2.2. IP-based telephony protocol

H.323, which is designed to operate above the transport layer of the underlying network, is popular as a set of protocols for Internet telephony, and H.323 terminals can talk to each other directly. An important part of H.323 is the H.245 call control protocol. When one H.323 terminal wants to call another, it uses H.245 to negotiate the properties of the call, and also H.245 can be used to signal the UDP port numbers that will be used by RTP and RTCP for the data stream in the call. Once this is accomplished, the call can proceed, with RTP being used to transport the data streams and RTCP carrying the relevant control information. In order to transmit speech signals over the simulated IP channel from the client to the remote recognizer in our work, two H.323 terminals are used to be local side and remote side, respectively.

In a general H.323 implementation, four logical entities are required: terminal, gateways, gatekeepers, and multipoint control units. It is possible to establish an H.323-enabled network with just terminals. For some kinds of applications, an MCU is required. The H.323 terminal provides real-time, two-way audio, video, or data communications with another H.323 terminal. A gateway is an optional component in an H.323-enabled network. The most common use of a gateway, which can provides data format translation, is to connect an H.323 network to the PSTN. A gatekeeper, which is optional, is the brain of an H.323 zone, which includes all the terminal gateways and multicast control units managed by one gatekeeper. Multipoint Control Units are used in the case of conferences with more than two users. They ensure that connections are properly set up and released.

2.3. packet loss issue

Transmission Control protocol (TCP) and UDP are used to find, access, and communicate with each other over IP [4]. TCP can compensate for any loss of packets by retransmitting packets that get lost. However, this introduces considerable delays, and is not fast enough for real-time communication. UDP is much faster and suitable for real-time transition of speech, but this cannot recover lost packets.

The routing of packets over the IP telephony network is controlled by routing protocols. Their function is to route the packets from their source to the destination over network. Nevertheless, when network traffic increases, it is possible that these packets get lost in this network.

For measuring the influence of missing speech packets on the ASR system performance, we use a Soekris net 4501 IP simulator made by the Engineering Soekris Engineering Company in order to control packet loss rate.

3. Enhancement algorithm

There exist a great number of techniques for enhancement of noise-corrupted speech, such as Ephraim-Malah MMSE, log spectral amplitude (LSA) estimation, nonlinear subtraction, configuration-based spectral estimation, speech enhancement based on human auditory perceptual criteria, etc. The purposes of speech enhancement are to improve the perceptual quality and reduce listener fatigue. Generally, the speech enhancement methods vary with various factors such as types of noise that they can cope with, complexity and their computational efficiency. Since our ASR is intended for IP use in diverse conditions, we focused our research work on enhancement schemes which are capable of dealing with stationary additive noise and the non-stationary case. Furthermore, these methods may be implemented in real time speech signal processing; therefore other considerations were taken as to the computational efficiency of the chosen enhancement methods. We trade off between computational efficiency and ASR performance taking into account the various possible scenarios of H.323 channel speech transmission; the MMSE-LSA method was chosen to examine the effects of speech enhancement that are capable of handling the possible non-stationary noise in the IP channel.

In particular the MMSE-LSA estimator is a short-time spectral amplitude method, which minimizes the mean-square error of the estimated logarithms of the spectra. Such a method is formulated as follows. An observed noisy speech signal y(t) is assumed to be a clean speech signal x(t) degraded by uncorrelated additive noise n(t). Thus

$$y(t)=x(t)+n(t). \tag{1}$$

Let $X_k=A_k e^{j\theta_k}, N_k$ and $Y_k=R_k e^{j\theta_k}$ denote the kth spectral component of the clean speech signal x(t), noise n(t) and the observed noisy speech y(t), respectively. We are looking for the estimate $\hat{A}_k$, which minimizes the following distortion measure

$$E\{(\log A_k-\log \hat{A}_k)^2\} \tag{2}$$

given the observed signal $\{y(t),0\le t\le T\}$. Obviously, the estimator is given by

$$\hat{A}_k=\exp\{E[\ln A_k \mid y(n)],0\le n\le N\}. \tag{3}$$

With the Gaussian model assumption, the gain function is given by

$$H(k)=\frac{\xi_k}{1+\xi_k}\exp\left(\frac{1}{2}\int_{v_k}^{\infty}\frac{e^{-t}}{t}dt\right) \tag{4}$$

where

$$\xi_k=\frac{E\{|X_k|^2\}}{E\{|N_k|^2\}},\quad \gamma_K=\frac{R^2{}_K}{E\{|N|^2\}},\quad v_k=\frac{\xi_k}{1+\xi_k}\gamma_k. \tag{5}$$

ξ_k and γ_k denote the a priori SNR and the a posteriori SNR, respectively. The estimate of a priori SNR, ξ_k, can be given by the decision-directed approach proposed in [1] and is described as follows

$$\hat{\xi}_k(l)=\alpha H_k{}^2(l-1)\gamma_k(l-1)+(1-\alpha)\max[\gamma_k(l-1),0],\ 0\le\alpha\le 1, \tag{6}$$

where l denotes the frame number.

4. Experiment platforms

4.1. Training and testing data corpus

To evaluate and compare the performance of speech recognition, the TIMIT and NTIMIT databases were adopted. The speech signals used for both training and testing in our recognition experiments were acquired from these speech databases. These corpuses contain a total of 6300 sentences, 10 sentences spoken by

each of 630 speakers from 8 major dialect regions of the United States. The speech data are stored in two different sets: train and test, and each set is then separated into 8 subsets: dr1 to dr8.

4.2. Noise signal processing

In this experiment, transmitted clean speech signals with a packet loss rate of 2% are mixed with several different kinds of noise sources [5].

In our experiment, we have chosen seven types of noise sources used previously in [6] for speech recognition investigation. In [6], a first and second moment analysis across four seconds of noise signal was conducted to determine the degree of stationarity for each noise source. A subjective score of stationarity, whose range is from one to ten, was assigned to each noise source based on this analysis, with a score of one denoting wide-sense stationary, and ten denoting non-stationary. Based on this criterion, the ten types of noise source were selected in following table.

Table 1 Description of noise sources. Stationarity denotes a measure of noise stationarity based on first and second moment analysis (1 refers wide-sense stationary and 10 refers non-stationary).

Noise	Stationarity	Description
WGN	1	White Gaussian noise
SUN	1	Cooling fan of SUN 4 noise
HEL	3	Helicopter fly-by noise
LCI	3	Large city noise
LCR	5	Large crowd noise
HWY	5	Noise in a car traveling
BAB	9	Noise-multiple speakers

The background noise d(n) was artificially mixed with the clean speech signals x(n) using

$$y(n) = x(n) + g.d(n), \quad (7)$$

where g is adjusted to achieve the desired signal-to-noise ratio (SNR):

$$SNR = 10\log\left(\frac{\sum_{n=1}^{N} x^2(n)}{\sum_{n=1}^{N} d^2(n)}\right), \quad (8)$$

where the summation is over the entire utterance. In order to have a consistent noise environment for comparison objective, the SNRs of 0, 5, 10, 15, and 20 dB were chosen to use in evaluation experiments.

4.3. ASR system implementation

Our baseline system is based on standard MFCC features. The speech signal is represented with an energy term to be appended to the 12 MFCCs, giving a 13-static parameter feature vector computed over a window of 25 ms with a frame rate of 100 Hz; the first and second derivatives are also added.

The phonetic acoustic models are composed of 3-state HMMs, which have no transition skips, and each state consisted of seven Gaussian mixtures with diagonal covariance matrices for the TIMIT database, and fifteen Gaussian mixtures with diagonal covariance matrices for the NTIMIT corpus. The training data use a decision tree-tying algorithm, and the standard Viterbi decoder is used in the recognition process. Also, a 3-state HMM is used to account for the silences surrounding each utterance, and a single state model for the inter-word short pauses. The speech recognition was based on the HMM paradigm as implemented by the HTK software. The speaker-independent tri-phone HMM-based baseline recognizer offered a word recognition correctness rate of 97.46% for the 260 test sentences in the test directory 2 of the TIMIT database. Similarly, we obtained a word recognition correctness rate of 90.75% for the NTIMIT database using the same structure ASR system.

5. Experimental results

In this work, H.323 protocols are used on top of a packet-based network transport to provide real-time point-to-point audio communication between two terminals, and it does not provide a guarantee that packets will be delivered at all. Due to time sensitivity of voice transmissions, packet losses greater than ten percent are generally intolerable. Consequently, in order to assess the influence of missing speech packets on the ASR system performance, we have considered several simulated random packet loss rates: 10, 5, 3, 2, and 1%. In the following Table 2, the test results are presented to be used as reference.

Table 2. ASR results over H.323 channel for the TIMIT database

Package loss rate %	Word correct rate %	Sentence correct rate %
0	97.46	83.46
1	95.84	76.15
2	93.07	66.15
3	92.72	63.85
5	88.60	51.15
10	82.46	33.46

Table 3. ASR results over an H.323 channel for the NTIMIT

Package loss rate %	Word correct rate %	Sentence correct rate %
0	90.75	69.62
1	86.63	53.85
2	83.46	44.23
3	80.45	35.77
5	76.94	33.85
10	65.06	15.35

Since complete segments of the signal are lost, packet loss severely affects the performance of a speech recognizer. Also we can conclude from Table 2 and Table 3 that ASR performance is not linearly deteriorated when packet loss increases linearly. Therefore, we cannot predict or evaluate the speech recognition performance from packet loss rate. In addition, Table 3 manifests that the ASR performance is reduced slightly since the NTIMIT database, which was used to simulate the PSTN telephone to another, the PSTN to PC, and PC to the PSTN telephone in our work, uses the telephone bandwidth.

Table 4. ASR Results in (%) word recognition correctness rate for the TIMIT corpus; speech corrupted by various types of additive noise.

Noise Signal	0 dB	5 dB	10 dB	15 dB	20 dB
WGN	3.42	10.17	31.35	53.84	75.98
SUN	5.11	14.20	44.28	74.53	85.36
HEL	8.94	28.23	55.24	78.17	89.43
LCI	7.63	24.55	61.33	81.41	88.56
LCR	8.04	22.05	50.28	77.95	88.16
HWY	29.39	51.17	69.09	83.12	88.43
BAB	10.51	17.19	39.41	70.14	79.57

Table 5. ASR Results in (%) word recognition correctness rate for the NTIMIT corpus; speech corrupted by various types of additive noise.

Noise Signal	0 dB	5 dB	10 dB	15 dB	20 dB
WGN	3.33	13.28	34.55	55.41	71.50
SUN	3.02	12.93	35.12	58.61	75.71
HEL	5.00	21.22	51.78	73.87	82.42
LCI	3.77	20.78	52.96	72.78	82.68
LCR	3.95	18.59	42.09	64.58	79.92
HWY	9.78	31.57	57.21	71.72	79.18
BAB	7.28	7.19	33.54	53.31	70.50

Table 6. ASR Results of the enhanced speech recognition using the MMSE LOG method in (%) word recognition correctness rate for the TIMIT corpus.

Noise Signal	0 dB	5 dB	10 dB	15 dB	20 dB
WGN	13.68	40.42	65.15	79.83	86.37
SUN	16.00	42.57	69.84	84.70	90.22
HEL	15.88	40.16	68.83	84.17.	90.44
LCI	14.69	37.13	64.36	83.38	90.71
LCR	12.34	32.92	63.52	83.56	91.28
HWY	70.85	82.33	88.82	92.06	92.24
BAB	5.44	9.19	24.51	56.42	76.37

Table 7. ASR Results of the enhanced speech recognition using the MMSE LOG method in (%) word recognition correctness rate for the NTIMIT corpus.

Noise Signal	0 dB	5 dB	10 dB	15 dB	20 dB
WGN	11.49	24.16	41.25	58.97	72.25
SUN	13.20	29.77	49.72	69.14	79.00
HEL	11.97	32.44	58.13	72.86	81.10
LCI	12.01	29.94	51.34	71.02	80.58
LCR	3.95	23.72	49.15	70.36	79.48
HWY	64.05	74.62	81.19	83.25	84.26
BAB	4.21	7.80	18.98	35.20	59.45

Now we evaluate the performance of the MMSE-LSA method. In the experiment, the noisy speech is obtained by corrupting the clean speech signals, which were transmitted over an H.323 channel with 2% packet loss rate, with white Gaussian noise, cooling fan noise, helicopter fly-by noise, city noise, large crowd noise, car noise and babble (multiple speakers) noise with different noise levels. As shown in Tables 4 to 7, this speech enhancement method leads to an apparent improvement in recognition performance for most of the noisy signals, excluding the BAB (speaker-like) noise. Because this kind of noise has a stationarity score of 9 as defined previously in [6], i.e., it is a highly non-stationary noise, the enhancement method fails to improve speech corrupted by it under all SNRs. More importantly, it can be clearly seen that at the same SNR the recognition rates vary considerably with the kinds of noises applied. As a result, SNR alone cannot be used to predict or evaluate the speech recognition performance. Based on experimental results, we conclude that this kind of enhancement scheme is suitably applied to slowly varying additive noises only, rather than quickly varying ones.

6. Conclusions

We have investigated the performance of the MMSE-LSA method under various experimental conditions: clean speech signals were transmitted via an H.323 channel under several specific packet loss rate conditions, then these speech data were corrupted by seven types of additive noise with different SNR. Based on our experiment, we conclude that since all the information concerning one or more frames is lost, these have a severe influence on recognition accuracy. Moreover, experiments have demonstrated that increased robustness to additive noise can be achieved by the MMSE-LSA front-end processing scheme. Future study will be continued on the packet loss in more detail and in the development of methods coping with packet losses. Also, further work will be focused on more effective speech enhancement techniques for improving the robustness in more different types of noisy environments.

7. References

[1] Y. Ephraim and D. Malah, “Speech enhancement using optimal nonlinear spectral amplitude estimator”, IEEE Trans. Acoust. Speech Signal Processing, vol ASSP-32, no. 6, pp. 1109-1121, 1984

[2] Y. Ephraim and D. Malah, “Speech enhancement using a minimum min-square error log-spectral amplitude estimator”, IEEE Trans. Acoust. Speech Signal Processing, vol ASSP-33, no. 2, pp. 443-445, 1985.

[3] ITU-T Recommendation H.323, “Packet-based multimedia communications systems”, Geneva, Switzerland, January, 1998.

[4] B.Douskalis, “IP Telephony: The integration of robust VoIP services” Hewlett-Packard, 2000.

[5] The noise data is available at http://www.ee.duke.edu/Research/Speech/rspl_software.html.

[6] J. H. L. Hansen and L. Arslan, “Robust feature-estimation and objective quality assessment for noisy speech recognition using the credit card corpus”, IEEE Trans. Speech Audio Process., vol. 3, no. 3, pp. 169-184, 1995.