ASR FAILURE PREDICTION BASED ON SIGNAL MEASURES

Lu Huo, Ulrich Heute

Institute for Circuit and System Theory, Faculty of Engineering,
University of Kiel, Germany

lhu@tf.uni-kiel.de

Abstract:

In a spoken-dialogue system, it is useful to know to which extent the automatic speech recognizer (ASR) can overcome the adversary environment of a real-world telephone system and still produce correct recognition results. This kind of information, from the system developer's point of view, can be used to adjust the recognition-result handling; and from the system-manager's point of view, can be used as an important system-evaluation factor. In this work, we study signal-based measures that influence the recognition performance, and we try to build a prediction model for the "Lucy" database.

1 Introduction

Commercial spoken-dialogue telephone services for tasks such as timetable or tariff information, reservation, or telephone banking are usually implemented on dialogue platforms, which possess speech-recognition and understanding, dialogue-management, speech-generation, and speech-output components [1]. The performance of the ASR, as the input-interface with the user, directly influences the usability of the whole dialogue system. As one integral part of a research project entitled "Testbed for Interactive Dialogue-System Evaluation (TIDE)" [2], we study signal measures that could predict ASR failure.

Two types of signal measures have been considered: One represents the channel distortions, and the other represents the user-signal characteristics.

Theoretically, any channel distortion that substantially changes the spectrum of the speech signal could negatively influence the ASR performance. So the impacts of codecs, simultaneous noise (background noise, circuit noise, multiplicative noise), packet loss, or channel attenuation should all be considered. In [3], the influence of noise and codecs on the ASR has been intensively studied; it was shown that, if the distortion is large enough, the performance of the ASR drops dramatically. In the literature, there are numerous reports on the development of robust algorithms for speech recognition against the deteriorating effect of packet loss on the ASR performance. In our work, channel attenuation, background noise level, multiplicative noise level, and signal-to-noise ratio have been estimated.

The user-signal characteristic is another important factor that may influence the ASR performance. Analysis of the user signals is much more complex than that of the channel distortion. In the case of a dialogue system, the users may use isolated words or continuous sentences; they may have a reading style or a spontaneous style in speaking; they may use a large vocabulary or restrict themselves to a few given words; they may have various emotions such as anger, fear, or hesitation in speaking. All these effects could go beyond the system capacity and cause ASR errors. How to automatically classify the user characteristics and model their influence on the ASR is still open to research. In [4], Hirschberg has found that prosodic features, such as the fundamental frequency (F0), the speech duration, and others, can be used to predict the recognition performance. Inspired by this work, statistics of F0 and speech duration are put under study here.

In section 2, the algorithms used to extract measures from the user's speech signal are described. The prediction model of the ASR failure, based on the database "Lucy", is given and discussed in section 3. In section 4, we discuss the difficulties we met in this work and ideas for possible solutions. Finally, a brief summary is given in section 5.

2 Signal Measures

The database "Lucy" was collected with a telephone-based system implemented on a Nuance dialogue platform. It generates log-files and records the user speech signal in each dialogue state. 25 native German test subjects interacted with a prototype implementation of this system. The subjects were recruited according to 5 groups: AF (adult female), AM (adult male), SF (senior female), SM (senior male), and C (child). All subjects had to carry out a minimum of 4 interactions with the system, targeting on wire-line telephone-tariff enquiry, mobile-telephone tariff inquiry, internet-tariff inquiry, and internet-problem report. This resulted in a set of 280 dialogues and 1672 user audio files in the database.

This database consists of data both on the signal level (degraded user signal) and on the symbolic level (log information). It contains also the usability evaluations of the users that is necessary for the usability prediction in TIDE. But as we will see later in this paper, it is not necessarily an ideal database for the study of the ASR performance.

In the following, there are brief descriptions of the signal measures of both the channel-distortion and user-signal characteristics.

2.1 Measures of the Channel Distortion

***Active Speech Level (ASL)*:** The ASL is calculated from active-speech segments extracted by the GSM Voice-Activity Detector (VAD) [5], following the algorithm described in ITU-T Rec. P.56 [6]. It consists of a histogram analysis with multiple variable thresholds, and it results in a level relative to the overload point of the digital system.

Noise Level: Noise mainly stems from background noise present at the user's side and picked up by the telephone handset, as well as from circuit noise induced by the subscriber line. Speech pauses have been extracted with the help of the GSM VAD [5], and then a smoothed noise-power spectrum is determined from the windowed non-speech segments.

***Signal-to-Noise Ratio (SNR)*:** With a similar processing as for the noise level, a smoothed power-spectral density of the speech signal is determined during speech activity. The SNR is calculated as the ratio between both powers, calculated per utterance.

***Mean Cepstral Deviation (MCD)*:** It is known that logarithmic-PCM or ADPCM coding introduce multiplicative noise in telephone channels. In order to determine the level of degradation introduced this way, we assume that the recorded speech signal *y(k)* is obtained by the addition of the clean speech signal *s(k)* and a white Gaussian noise component *n(k)* with a certain ratio *Q*:

$$y(k) = s(k) + s(k) \cdot 10^{-Q/20} \cdot n(k) \qquad (1)$$

Falk et al. [8] proposed to measure this noise via the flatness of the output-speech signal *y(n)*. The underlying idea is that – because the multiplicative noise of Eq. (1) introduces a fairly flat noise in the spectral domain – the lower the *Q* value, the less the spectrum of *s(n)* can be preserved in *y(n)*, and the flatter is the spectrum of *y(n)*.

We use the MCD as a measure of the amount of multiplicative noise present in the degraded speech signal, because analyses have shown that the correlation between *Q* and MCD is about

-0.93 [8]. To calculate MCD, we computed cepstral coefficients for the active-speech frames (decided by the GSM VAD), and average their standard deviations.

Single-Ended Speech Quality Estimate: Multiplicative noise is not the only degradation introduced by modern telephone channels. In particular, non-waveform speech codecs generate distortions which have different perceptual and signal correlates, and which have shown to degrade recognition performance [1]. In order to cover these channel degradations, we use the single-ended model described in ITU-T Rec. P.563 [6] to obtain an indication of the overall speech-quality degradation introduced by the channel. This model generates a clean-speech reference from the degraded speech signal by means of an LPC analysis and re-synthesis. Both the generated clean and the recorded degraded speech signals are transformed to a perceptually motivated representation. An estimate of the mean-opinion score (MOS) is then determined from a comparison of both representations.

2.2 Measures of the User-signal Characteristics

***Active Speech Duration (AD)*:** We use the GSM VAD to cut off pauses at the beginning and at the end of the speech signals that remain after the VAD of the dialogue platform.

***Fundamental Frequency (F0)*:** Hirschberg et al. [4] have shown that mean and maximum F0 can be useful in predicting the recognition error. We adopted the autocorrelation analysis from Rabiner [9] and some simple smoothing algorithm for F0 estimation. For each user utterance, the mean, the standard deviation, and the 95% percentile of F0 are calculated.

2.3 Measurement Result Analysis

Table 1 summarizes the above-mentioned measures for the database “Lucy”. It shows that the utterances have a high SNR and are generally of a relatively high speech quality. This could be due to the test set-up where subjects interacted with the system from two test cabinets equipped with a good wireline telephone. In addition, the utterances are relatively short, indicating that the subjects preferred to use a simple command language towards the system.

Parameter	Mean	STD
ASL (dB)	-24.5	7.8
Noise level (dB)	-57.4	7.9
SNR (dB)	32.1	13.3
MCD	0.103	0.008
Single-ended estimate	2.46	0.85
Active speech duration (s)	1.53	1.87
F0_MEAN(Hz)	165.0	45.2
F0_STD (Hz)	38.7191	25.7884
F0_95PERCENTILE (Hz)	220.5846	73.4591

Table 1 - Summary of measures from database „Lucy“

3 Prediction of ASR Performance

In order to represent recognizer performance, the recognition results are classified into two classes, namely, “correctly recognized” and “error”; the first term refers to the complete match between the transcript of the user signal and the recognition result, while the second refers to any mismatch or rejected situation. **Table 2** summarizes the ASR performance, separated for the different user groups.

We see that the “Lucy” database has, on one side, a large complexity in the speaker domain and task domain, which could have greatly influenced the recognition result (**Table 2**). On the other side, it has low detectable channel distortions (Table 1). This adds to the difficulties in analyzing the significance of the measures.

Person Type	No. Recognition	No. Rejected	No. Correct	Correct%
Adult male	295	41	273	81.25
Adult female	248	52	225	75.00
Senior male	193	82	166	60.36
Child	131	63	119	61.34
Senior female	97	92	85	44.97
All	964	330	868	67.08
All but senior female	867	238	783	70.86

Table 2 - Summary of the ASR performance among the different user groups

The user group SF experienced most of the failure, because they tended to use complex and long sentence; so, we have deleted this user group from the database to be analyzed. After deletion, 29.14% of the 1105 remaining recognition results are classified as “error”. So without any parameters in the prediction model, we can predict a recognition failure with an error rate of 29.14%. Any model that reduces this error rate can help to predict the recognition performance more reliably.

Inspired by the work of Hirschberg [4], we also adopt the rule-learning program “RIPPER” from Cohen [10]. “RIPPER” is used here to generate plausible rules based on the extracted parameters from the speech signal that can be used to predict the ASR performance. Fig. 1 is an example of the resulting rules from “RIPPER” that can decrease the prediction error from the baseline 29.14 % to 21.63%

```
if AD>=1.39 && MCD>=0.11, then Error.
if AD>=1.82 && SNR>=-51.30 && MCD>=0.098 && MCD<=0.11& F0_MEAN>=129.92, then Error.
if AD>=0.92 && F0_MEAN<=105.14, then Error.
Else Correct.
```

Figure 1 - Rule set for predicting recognition errors; its prediction-error rate is 21.63%+1.24%

The parameters used in the prediction model are selected using the “stepwise” method. At first, we just select one of the nine candidate parameters into the prediction. Then more parameters are added into the model one by one; one parameter would be then selected only if it could decrease the prediction error. Table 3 shows the selected best results for the prediction model with a number of **1, 2,** 3, 4, and 5 parameters, respectively.

Parameters included	Error %	Parameters included	Error %
AD+MCD+NL+ASL+SNR+F0_MEAN	20.69	AD+MCD or AD+ASL	22.71
AD+MCD+SNR+ F0_MEAN	21.18	AD	25.52
AD+MCD+ASL	22.26		

Table 3 - Prediction error for predicting the ASR performance

From these results it seems that speech duration and multiplicative noise are the most significant parameters in the prediction model using our database.

4 Discussion

In the TIDE project, our analyses were limited by the effects contained in our data collection. A more suitable database should possess variance in a large set of factors that influence the dialogue system. Here we want to address some issues of such a suitable database just from the point of view of ASR performance evaluation.

4.1 Test Conditions

We have found that in our database the speech duration (AD) and multiplicative noise (MCD) are the most significant factors. But this result is far from a conclusion about the general situation in the ASR of telephone-based dialogue services, because, in "LUCY", there is a lack of some distortions existing commonly in modern systems, such as background noise, packet loss / frame loss. A more realistic telephone environment is desirable for studying the influences of channel distortions on the recognition performance.

4.2 Recording Point

Another problem with "Lucy" is that the speech signals are only segments of the user signals after VAD. If, for some reason, the VAD could not decide between speech activity and non-speech activity correctly, then either the non-speech activity would be selected for further recognition or the useful speech signal would be cut off. Both situations would directly result in recognition errors. This kind of errors cannot be analyzed using recordings after VAD. Another drawback of such short segments is that, due to short duration, it is difficult to obtain reliable signal measures, because some measurements require a signal length of at least 3 to 4 seconds. For a complete and reliable analysis of the recognition failure, the speech data should be recorded before VAD.

4.3 Database Scenario for Evaluation of a Telephone-Based Dialogue System

In "Lucy", we have only the degraded user signal, as received by the dialogue system. This leads to difficulties in detecting and measuring some distortions such as echo, which is already manually detected from the database. So we need to discuss here what kind of recording scenario may be suitable for a test-bed for a telephone-based dialogue system.

Generally, the databases can be classified by three different scenarios (see Fig. 2):

1) In an intrusive scenario, one compares the input and output of the transmission channel. So records at points A and B are necessary for evaluating the user signal and records at points C and D for evaluating the system utterance.
2) In an INMD (in-service non-intrusive measurement device) scenario, one compares the near-end and the far-end signals for each terminal, i.e., records at points B and D are necessary for evaluating the user signal, and audio signals A and C for evaluating the system utterance.
3) In a single-ended scenario, only the degraded signals at the output of the transmission channel are available. We need the audio signals at points B and C to study the user signal and the system utterance respectively. But using this scenario seems too difficult to detect some distortions such as echoes.

In our test-bed of telephone-based dialogue systems, both the user utterance and system utterance influence the user's opinion of the system. So, ideally, we should record at all the four record points. But in the context of system management, only records at points B and D are practically recordable by systems, i.e., only an INMD scenario is realizable. Using records at D, the quality of the system utterance (probably a Text-to-Speech output) could also be evaluated in this scenario with a suitable algorithm.

Figure 2 - Telephone-based dialogue system

5 Summary

In this work, we have studied ASR failure based on signal measures. Certain measures, such as speech duration (AD) and multiplicative noise (MCD), are found significant in the database "Lucy". In the discussion we have addressed the issues of a more suitable database for the quality evaluation of ASR performance in telephone-based dialogue systems, which may be useful in preparation of future-experiment designs for a dialogue-system evaluation.

6 Acknowledgements

The described work was supported by the TIDE project funded by Deutsche Telekom AG. The authors would like to thank the colleagues in Deutsche Telekom Laboratories (T-Labs), Forschungszentrum Telekommunikation Wien (FTW), and T-Systems Enterprise Services GmbH (T-Systems) for their cooperation in this project.

Literature

[1] Möller, S., Quality of Telephone-Based Spoken Dialogue Systems, Springer, NY, 2005.
[2] Möller, S., Engelbrecht, K.-P., Pucher, M., Fröhlich, P., Huo, L., Heute, U., Oberle, F.: "TIDE: A Testbed for Interactive Spoken-Dialogue System Evaluation", accepted in 12th Int. Conf. "Speech and Computer" (SPECOM'2007), October 15-18, Moscow, 2007.
[3] Möller, S., Kavallieratou, E., "Diagnostic Assessment of Telephone Transmission Impact on ASR Performance and Human-to-Human Speech Quality", in: Proc. 3rd Int. Conf. on Language Resources and Evaluation (LREC 2002), Vol. 4, 1177-1184, 2002.
[4] Hirschberg, j., Litman, D., Swerts, M., "Prosodic and other cues to speech recognition failures", Speech Communication 43, 2004.
[5] ETSI ETS 300 040, European Digital Cellular Telecommunications System (Phase 1); Voice Activity Detection (GSM 06.32), ETSI, Sophia Antipolis, 1992.
[6] ITU-T Rec. P.563, Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications, ITU-T, Geneva, 2004.
[7] ITU-T Rec. P.56, Objective Measurement of Active Speech Level, ITU, Geneva, 1993.
[8] Falk, T.H., Chan, W.-Y., "Single-Ended Speech Quality Measurement Using Machine Learning Methods", IEEE Trans. Audio Speech Language Proc., 14(6):1935-1947, 2000.
[9] Rabiner, L., "On the Use of Autocorrelation Analysis for Pitch Detection", IEEE Trans. Acoustics, Speech and Signal Process. 25(1):24-33, 1977.
[10] Cohen, W., "Learning trees and rules with set-valued features", 14th Conference of the American Association of Artificial Intelligence (AAAI), Portland, pp.709-716.