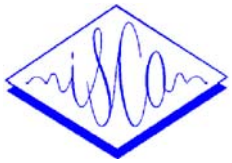

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

Robust Speech Recognition for Wireless Networks and Mobile Telephony

Reinhold Haeb-Umbach

Philips GmbH Forschungslaboratorien
P.O. Box 50 01 45
D-52085 Aachen, Germany
Email: haeb@pfa.research.philips.com

ABSTRACT

The increased popularity of mobile telephony introduces both challenges and opportunitites for automatic speech recognition. ASR offers ways to simplify the use of mobile phones, notably in hands- and eyes-busy situations. However, the acoustic environment can be severely degraded and the wireless network may add additional distortions to the speech signal. This paper gives an overview of the sources of degradation and attempts to robust speech recognition for mobile communications. Emphasis is placed on approaches which are suitable for implementation in mobile terminals. Two example applications are described which illustrate the robustness issues and design considerations typical of low-cost noisy speech recognition: voice-dialling in a GSM phone and hands-free digit recognition in the car.

1 INTRODUCTION

Speech recognition is considered a viable approach to simplify the use of small and highly sophisticated mobile phones. Voice dialling is currently probably the most popular example application. Hands-free voice control is considered particularly useful for mobile terminals in the car, since it can eliminate the safety hazard introduced by manually operating a car telephone, radio or navigation system. The importance of in-car speech recognition is reflected by the existence of publicly funded projects such as VODIS [1], supported by the Commission of the European Union, and MoTiV [2], which is partly funded by the German Ministry of Education and Research. Among others, an important objective of these projects is the collection of realistic speech databases to foster research and application development for in-car speech recognition. Much insight in the nature of the problem has been gained from the predecessor ESPRIT-ARS project. Within the European SpeechDat project 1000-speaker speech databases are recorded from the GSM network in 5 languages [3].

Since mobile phones can be operated virtually anywhere, the speech signal at the input of the recognizer can be severely degraded. Distortions are caused by ambient noise sources (e.g. if calls are placed from public places, at the roadside, hands-free from within a car, etc.). Note that the distortions are often nonstationary. Further degradations are due to the characteristics of the signalling in wireless networks (speech coding loss, bit errors, fading, dropouts, etc.).

It is difficult and expensive to capture all possible signal degradations in a training database. Recognition techniques are required which are robust to whichever environmental conditions are present. Many approaches to robust speech recognition have been proposed in the literature, see e.g. [4-7] for tutorials. In this paper we focus on those methods which can be implemented in mobile terminals. Here, hardware constraints and real-time performance are important issues. We describe two example applications of mobile terminal-based recognizers which demonstrate the robustness issues and design considerations typical of low-cost noisy speech recognition:

- Voice dialling in a GSM phone, and
- Hands-free digit recognition in the car.

The first is an example of speaker-dependent recognition with typically large mismatch between training and recognition. We describe robust feature transformations, give typical error rates measured in the field and discuss implementation issues.

The second is a considerably more difficult task. We will review single-microphone noise reduction techniques and present experimental results on data recorded in a driving car.

The organization of this paper is as follows. In the next section we discuss specific aspects of network-based versus terminal-based speech recognition. Section 3 gives an overview of approaches to robust speech recognition in mobile terminals, and Sections 4 and 5 describe the forementioned examples, voice dialling in a GSM phone and digit recognition in the car, respectively.

2 NETWORK- VS TERMINAL-BASED RECOGNITION

There is the principle choice as to whether the speech recognizer is physically located in the network or in the terminal.

The advantage of a recognizer in the network is that the pressure on the hardware costs is less pronounced since the same hardware is used for many subscribers. This allows for more sophisticated recognition algorithms. Applications which demand high-performance processors and large RAM are possible only with a network-hosted recognizer (e.g. spoken dialogue systems, remote dictation). Approaches not feasible in mobile terminals may turn out to be advantageous, such as the implementation of voice dialling with speaker-independent subword units [8]. While a network-based server can be easier maintained and upgraded, a potential disadvantage is the

degradation of the speech signal by the transmission over the network. Studies have been conducted to quantify the influence of speech coding algorithms on automatic speech recognition [9-11]. It was reported that when the coding rate of the training data was 64 kb/s (PCM), a reduction of the bit rate of the test data to 13 kb/s (GSM fullrate) led to an increase in error rate by up to a factor of two [9]. However, a great deal of the degradation could be recovered by using the same coding scheme in training and recognition. GSM halfrate (5.6 kb/s) performs similar to GSM fullrate, execpt in noisy environments where a significant loss in performance was observed with the halfrate coder [10].

In order to overcome the network channel limitations, distributed speech recognition can be employed: the speech recognition front-end (i.e. the computationally simple feature extraction module) is located in the handset, and the feature stream is transmitted to the network-based powerful multi-user recognition server. The output of the recognition can be passed back to the user for further processing, if necessary. This approach, however, requires a standardized feature extraction function, which is currently not in place.

A speech recognizer in the handheld does not suffer from network distortions. However, due to the limited hardware resources, only low-complex recognition tasks can be carried out, such as voice dialling or voice control. Voice control is preferably implemented in the terminal since this allows a better integration in the overall user interface.

While first terminal-based voice dialling systems had been introduced already a few years ago, e.g. [12], the advances in VLSI have only recently made possible a realization of the feature at virtually zero additional hardware costs.

3 ROBUST SPEECH RECOGNITION IN MOBILE TERMINALS

Approaches to improve the robustness of a speech recognizer basically fall in two broad categories.

The first category comprises techniques to reduce the sensitivity of 'characteristics' of the speech signal as computed in the recognizer towards possible signal distortions. These 'characteristics' can be features or distances or likelihoods. In this category we classify speech enhancement techniques, which remove the effect of noise in a preprocessing stage, robust feature extraction techniques, such as PLP, and feature transformation methods, such as subband filtering or the inclusion of delta features. All these methods aim at finding a representation of the speech signal which is less sensitive to distortions. Later in the recognizer distances or likelihoods are computed between observations and training references. Research has been conducted to devise distortion measures which are less sensitive to noisy test data.

In the second category we classify adaptation and compensation techniques which try to reduce the impact of the mismatch between training and test environments by either modifying the noisy feature vectors or by modifying the models such that the test feature vectors and the trained models are "closer". The compensation can be estimated from stereo data, from a parametric model of the compensation or as a byproduct during recognition from aligning the noisy test data with the trained models.

Practical constraints often limit the choice of measures to be used to improve robustness. When implemented in a mobile terminal, the following considerations must be taken into account:

- The technique should make as little assumptions on the type of distortions as possible. It has been outlined earlier that there is virtually no control over the signal quality, since mobile systems by definition can be used in almost any environment.
- The technique should be suitable for real-time processing: the user expects the recognition result right after he has uttered the command.
- The technique should require as little computing and memory resources as possible, since hardware costs, size, and power drain are important issues for mobile terminals.

These constraints are more likely to be fulfilled by measures of the first category, as we will see in the next section.

4 VOICE DIALLING IN A GSM TELEPHONE

Voice dialling is a speaker-dependent isolated word recognition task with typically a large mismatch between training and recognition. Training is carried out in a quiet environment while recognition often takes place under noisy conditions.

An approach to reduce the influence of both additive and convolutional noise is to place more emphasis on features characterizing the temporal dynamics of the signal. The first derivative has been shown to be not only effective for recognition of clean speech but also for Lombard and noisy Lombard speech [13] and for channel distorted speech [14]. If there has to be made a choice between static and delta features, delta features should clearly be preferred in mismatch situations.

Since delta features may be viewed as high-pass filtered static features, there are close links to subband filtering approaches such as RASTA [15] or cepstral (log-spectral) mean subtraction. The latter methods have been proposed to reduce the influence of unknown linear filtering. While the transfer characteristics may vary in detail, they all have in common a spectral null at dc of the modulation spectrum. The subband filters usually exhibit sharper nulls than the delta features. We have chosen a feature vector of static features only with a subsequent high-pass filter with a time constant similar to the delta features to have both the advantages of delta features and subband filtering with a low-dimensional feature vector [16].

A robustness issue, which is particularly important from a user satisfaction point of view, is the reliability of the recognition result. If the reliability of the hypothesized word were known, this could be exploited in the interaction with the human, e.g. by asking for repetition if a recognized word is considered unreliable rather than interpreting a misrecognized command. A deletion error of the recognizer is less annoying than a substitution or insertion error since a repair for the latter error types can be more cumbersome. Another

application of a reliability measure is the detection of out-of-vocabulary utterances.

Reliability or rejection decisions are usually derived from a likelihood ratio test which is a statistical hypothesis test to determine whether or not a sequence of feature vectors was generated by the model of a hypothesized vocabulary word or by a model corresponding to the alternate hypothesis. The likelihood for the alternate hypothesis may either be obtained from the acoustic models of the recognition vocabulary or from explicit "garbage" or "antikeyword" models [17,18], which have to be trained as separate models in addition to the models of the recognition vocabulary.

Table 1 presents recognition and rejection results for voice dialling, where training was done in an office and the test data had been generated by adding noise samples recorded in typical operating environments (car, public place, roadside) at different gain mixing ratios. The recognition vocabulary consisted of 10 words. Each of 50 speakers spoke two utterances of each word for training and 6 utterances for testing. In addition each speaker uttered 10 words not belonging to the recognition vocabulary to test out-of-vocabulary rejection. Note that the percent substitutions and deletions are measured on the utterances of in-vocabulary words, while the percent false alarms are measured on the utterances of the out-of-vocabulary words.

Gain mixing ratio	%subst.	%deletions	%false alarm
∞ (office)	0.8	6.3	22.3
30	0.7	6.6	24.3
25	0.8	6.9	23.3
20	0.6	7.8	24.7
15	0.4	9.0	21.2
10	0.5	12.0	18.8
5	0.4	22.0	12.8

Table 1: Performance of voice dialling system (average over 50 speakers)[1].

For rejection we employed a combination of two criteria which do not require explicit rejection models: First, the ratio of the likelihood of the best to the second best word hypothesis must exceed a certain threshold. Second, the ratio between the likelihood of the best word hypothesis and an alternate hypothesis computed from a one-state model, whose mixture emission probability consists of the sum of all component densities of all model states, must exceed a second threshold.

It is interesting to note from the above table that the substitution rate is virtually independent of the signal-to-noise ratio. With increasing noise level, the number of deletions, however, increase. This was considered less annoying than a corresponding increase in substition rate.

[1] The author is grateful to Thomas Eisele who has done this investigation.

5 HANDS-FREE DIGIT RECOGNITION IN THE CAR

With the spread of hands-free equipment the acoustical phonemona in the car environment have become an active field of research. Three basic factors have been investigated: the noise power spectral density, reverberation effects, and echo phenomena.

The background acoustic noise field in the car is dominated by low frequency components (< 500 Hz), caused by motor noise and road surface noise. At higher speeds noise introduced by airflow results in an increase in noise power for frequencies above 1 kHz [19]. In hands-free situations the overall noise power can reach very high levels so that signal-to-noise ratios below 5 dB are not unusual. Note that with such low SNRs the speaker intuitively adjusts his way of speaking to the ambient noise (Lombard effect).

The acoustic impulse response of the mouth-to-microphone channel has been measured for different microphone positions [20]. Typical values of the reverberation time T_{60} have been shown to be on the order of 50ms or below. These values are not critical for a human ear but could be a potential problem for an automatic speech recognizer.

The third phenomenon studied is the acoustic coupling between the loudspeaker and the microphone. This is more important for hands-free communication than for recognition, since the far end speaker can be annoyed by hearing his echo in full-duplex communication. The echo delay is mainly determined by the transmission network and is close to 200ms in the GSM network [20].

Noise reduction techniques, i.e. measures to suppress the uncorrelated additive noise, have been extensively studied. Today, most practical interest is still in single-microphone systems. Noise reduction methods differ in the way they estimate the short-term spectral amplitude of speech [21]. For speech recognition nonlinear spectral subtraction has been reported to be particularly useful [22].

It is important to note that for a hands-free communication system in the car the target recognizer can be both a human and a machine. In mobile telephony, call setup can be carried out by giving speech commands to a recognizer, and, once the call is established, speech is transmitted to the human at the other end of the line. Therefore there is considerable interest in speech enhancement techniques which improve performance for both types of "communication".

However, it is well known that techniques that lead to the best subjective speech quality improvement do not necessarily deliver the best automatic speech recognition performance. In an automatic speech recognizer recognition performance is worse if training and test environments do not match compared to the case of matched conditions — even if in the mismatch case the test data exhibit less noise than in the matched case. To achieve matched conditions it is therefore legal and common practice to artificially distort the test signal by measures such as noise masking [23] or SNR normalization [24].

Table 2 presents word error rates for speaker-independent connected-digit recognition in the car[2]. While non-linear spectral subtraction improves the error rate already by almost 20%, additional improvement is gained by normalizing the data to some target SNR, i.e. by a controlled degradation of the signal. See [25] for details of the experiments and the database.

	Baseline	NSS	NSS + SNR normal.
WER [%]	17.3	14.7	12.0

Table 2: Word error rates (WER) for nonlinear spectral subtraction (NSS) and the combination of NSS with SNR normalization (target SNR: 11dB).

6 CONCLUSIONS

The two example applications presented in this paper, voice dialling in a GSM phone and connected-digit recognition in the car, illustrated the robustness issues and design considerations typical of low-cost speech recognition in mobile terminals. Today first systems are in the market place. However, it will remain an active area of research to devise systems which achieve a good compromise between reliable recognition, ease-of-use of the device, and additional hardware costs.

7 REFERENCES

[1] URL: http://www.ic.cmu.edu/VODIS

[2] URL:http://www.tuev-rheinland.de/tsu/lfUE/pt_bvt/motiv/haupts.htm.

[3] H. Hoege, H.S. Tropf, R. Winski, H. v.d. Heuvel, R. Haeb-Umbach and K. Choukri, "European Speech Databases for Telephone Applications", in *Proc. ICASSP-97*, Munich, pp 1771-1774, 1997. See also URL: http://www.phonetik.uni-muenchen.de/SpeechDat.html.

[4] B. H. Juang, "Speech Recognition in Adverse Environments", *Computer, Speech and Language*, No. 5 pp 275-294, 1991.

[5] Y. Gong, "Speech Recognition in Noisy Environments: A Survey", *Speech Communication*, Vol. 16, pp 261-291, 1995.

[6] J.C. Junqua and J.P. Haton, "Robustness in Automatic Speech Recognition", Kluwer Academic Publishers, 1996.

[7] *Proc. ESCA-NATO workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, 1997.

[8] R.C. Rose, E. Lleida, G.W. Erhart and R.V. Grubbe, "A User Configurable System for Voice Label Recognition", In *Proc. ICSLP-96*, Phiadelphia, pp 582-585, 1996.

[9] S. Euler and J. Zinke, "The Influence of Speech Coding Algorithms on Automatic Speech Recognition", in *Proc. ICASSP-94*, pp I621-I624, Adelaide, Australia, 1994.

[10] S. Dufour, C. Glorion and P. Lockwood, "Evaluation of the Root-Normalised Front-End (RN_LFCC) for Speech Recognition in Wireless GSM Network Environments", in *Proc. ICASSP-96*, Atlanta, pp 77-80, 1996.

[11] B.T. Lilly and K.K. Paliwal, "Effect of Speech Coders on Speech Recognition Performance", in *Proc. ICSLP-96*, Philadelphia, pp 2344-2347, 1996.

[12] S. Dobler, D. Geller, R. Haeb-Umbach, P. Meyer, H. Ney and H.W. Rühl, "Design and Use of Speech Recognition Algorithms for a Mobile Radio Telephone", *Speech Communication*, Vol. 12, pp 221-229, 1993.

[13] B. Hanson and T. Applebaum, "Robust Speaker-Independent Word Recognition Using Static, Dynamic and Acceleration Features: Experiments with Lombard and Noisy Speech", in *Proc. ICASSP-90*, pp 857-860, 1990.

[14] J.C. Junqua, S. Valente, D. Fohr and J.F. Mari, "An N-Best Strategy, Dynamic Grammars and Selectiveley Trained Neural Networks for Real-time Recognition of Continuously Spelled Names over the Telephone", in *Proc. ICASSP-95*, Detroit, pp 852-855, 1995.

[15] H. Hermansky, A. Bayya, N. Morgan and P. Kohn, "Compensation for the Effect of the Communication Channel in Auditory-Like Analysis of Speech (RASTA-PLP), in *Proc. Eurospeech*, Genova, pp 1367-1370, 1991.

[16] H.G. Hirsch, P. Meyer and H.W. Ruehl, "Improved Speech Recognition Using High-Pass Filtering of Subband Envelopes", in *Proc. EUROSPEECH*, Genova, pp 413-416, 1991.

[17] J.G. Wilpon, L.R. Rabiner, C.H. Lee and E.R. Goldman, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 38, No. 11, pp 1870-1878, Nov. 1990.

[18] R.C. Rose, B.H. Juang and C.H. Lee, "A Training Procedure for Verifying String Hypotheses in Continuous Speech Recognition", in *Proc. ICASSP-95*, Detroit, pp 281-284, 1995.

[19] M.M. Goulding and J.S. Bird, "Speech Enhancement for Mobile Telephony", *IEEE Trans. Vehicular Technology*, Vol. 39, No. 4, pp 316-326. Apr. 1990.

[20] P. Scalart and A. Benamar, "A System for Speech Enhancement in theContext of hands-Free Radiotelephony with Combined Noise Reduction and Acoustic Echo Cancellation", *Speech Communication*, Vol. 20, pp 203 - 214, 1996.

[21] P. Scalart and J. Filho, "Speech Enhancement Based on A Priori Signal To Noise Ratio Estimation", in *Proc. ICASSP-96*, Atlanta, pp 629-637, 1996.

[22] P. Lockwood and J. Boudy, "Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the Projection, for Robust Speech Recognition in Cars", *Speech Communication*, Vol. 11, pp 215-228, 1992.

[23] A. Varga and K. Ponting, "Control Experiments on Noise Compensation in Hidden Markov Model Based Continuous Word Recognition", in *Proc. EUROSPEECH-89*, Paris, France, pp 167-170, Sep. 1989.

[24] T. Claes and D. Van Campernolle, "SNR-Normalisation for Robust Speech Recognition", in *Proc. ICASSP-96*, Atlanta, pp 331-334, 1996.

[25] D. Langmann, A. Fischer, F. Wuppermann, R. Haeb-Umbach and T. Eisele,"Acoustic Front Ends for Speaker-Independent Digit Recognition in Car Environments", in *Proc. EUROSPEECH-97*, Rhodes, 1997.

[2] The fairly high baseline error rate is due to the difficult nature of the data and the small amount of training data. We ran a control experiment on NOISEX and achieved state-of-the-art recognition rates [25].