

Noisy Speech Recognition Failure Diagnosis Using Minimum Message Length Decision Trees

Milos Cernak and Sakhia Darjaa

Institute of Informatics

Slovak Academy of Sciences

Complete Address: Dubravska Cesta 9, Bratislava, 084 07, Slovakia

Phone: (421) 2-5941 1129 Fax: (421) 2-5477 1004 E-mail: Milos.Cernak@savba.sk

Keywords: Automatic speech recognition, failure diagnosis, machine learning, decision trees

Abstract – Current ASR technology lacks of effective failure diagnosis of ASR systems. Figures of merits such as WER are very useful, but don't bring much insight into error patterns, error predictions or error analysis of ASR output. This paper explores an application of Minimum Message Length (MML) style decision trees for such a diagnosis, focusing on theoretical background and the failure diagnosis of noisy speech recognition. In addition, the paper focuses on failure diagnosis of noisy speech, covering several kinds of intrinsic speech variabilities as well. Results on added speech-shaped noise at different SNR, ranging from 25dB to -10dB, are presented.

1. INTRODUCTION

Our recent unpublished work on failure diagnosis using decision trees showed good performance of MML style decision trees for a task of speech recognition diagnosis. The work introduced minimum encoding style decision trees for the use in the topic of the evaluation of human and computer system performance. In addition, a novel approach in rating the reliability of this diagnosis using confusion matrices was presented. The performance of minimum encoding style decision trees for failure diagnosis was comparable to CART method [1], and they both outperformed C4.5 method [2].

The previous work of Douglas *et al.* considered the performance of speech recognition in noise and focused on its sensitivity to the acoustic feature set [3]. The authors used the decision tree CART method for data analysis, and they found, that the most significant factors related to human digit misrecognition were the speaker and the listener. The authors observed that once again the speaker is the most significant factor partitioning the data on the basis of digit recognition rate.

We present in this paper the analysis of computer speech misrecognition. Setup of ASR system is the same as used in [4]. Recognition of logatoms (see Section 3.1) bearing several intrinsic and extrinsic speech variabilities were analyzed using minimum encoding style decision trees. The theory of the approach is introduced in Section 2. Section 3 describes the experiments, and Section 4 discusses the results and future work.

2. MINIMUM ENCODING STYLE TREES

Minimum encoding approaches were developed for “fitting models to data” problems. The problem of finding a good model is converted to a problem of finding minimum encoding of the data, using concepts from Shannon’s theory of information.

Let us briefly introduce the probabilistic framework. Failure diagnosis task is based on learning, where we usually consider belief about hypothesis \mathbf{H} . Let H denotes a set of hypotheses. Then, for some $\mathbf{H} \in H$, $\Pr(\mathbf{H} | E_N)$ denotes the measure of belief that the hypothesis \mathbf{H} is “true” conditioned on the evidence $E_N = (e_1, \dots, e_N)$, a set of classified examples.

2.1 Minimum message length (MML)

Wallace and co-workers adopts a method they refer to as *Minimum message length (MML)* [5]. Given evidence E , the approach finds a hypothesis \mathbf{H} and additional parameters θ that maximizes the quantity.

$$\frac{\Pr(\mathbf{H}, \theta) \Pr(E | \mathbf{H}, \theta)}{\sqrt{I_E(\mathbf{H}, \theta)}}, \quad (1)$$

where $I_E(\mathbf{H}, \theta)$ is the *Fisher information* for evidence E . The Fisher information is a way of measuring the amount of information that the observable random variable \mathbf{H} carries about an unknown parameter θ upon which the likelihood function of θ , $\Pr(\mathbf{H}, \theta)$, depends. The Fisher information can be written as

$$I(\mathbf{H}, \theta) = \mathbb{E} \left\{ \left[\frac{\partial}{\partial \theta} \ln(\Pr(\mathbf{H}, \theta)) \right]^2 | \theta \right\}. \quad (2)$$

If the following regularity condition is met:

$$\int \frac{\partial^2}{\partial \theta^2} \Pr(\mathbf{H}, \theta) dx = 0, \quad (3)$$

then the Fisher information may also be written as:

$$I(\mathbf{H}, \theta) = \mathbb{E} \left[\frac{\partial^2}{\partial \theta^2} \ln(\Pr(\mathbf{H}, \theta)) | \theta \right]. \quad (4)$$

Thus Fisher information is the negative of the expectation of the second derivative of the log of $\Pr(\cdot)$ with respect to θ . Then, the Fisher information for evidence E , $I_E(\mathbf{H}, \theta)$ is given by [7, p57]:

$$\det\left(-\mathbb{E}_{E \sim \mathbf{H}, \theta}\left(\frac{D^2}{D\sigma D\sigma} \log \Pr(E | \mathbf{H}, \theta)\right)\right), \quad (5)$$

where $\det(\cdot)$ corresponds to the standard matrix determinant, σ are the $|\mathbf{H}| + |\theta|$ continuous parameters in the model for \mathbf{H}, θ , and $\frac{D^2}{D\sigma D\sigma}$ is the matrix of second derivatives.

From Shannon's theory of information we know that in an optimal code, the message length of an event with a probability P is given by negative binary logarithm of the probability P . Taking this into account and the quantity expressed by Equation (1), Wallace interprets Equation (5) that we wish to minimize the *code-length* (message length in the Shannon's terminology) for evidence E of

$$-\log \Pr(\mathbf{H}, \theta) - \log \Pr(E | \mathbf{H}, \theta) + \frac{1}{2} \log I_E(\mathbf{H}, \theta), \quad (6)$$

where $\frac{1}{2} \log I_E(\mathbf{H}, \theta)$ measures the encoding of the precision of the hypothesis, $-\log \Pr(\mathbf{H}, \theta)$ measures of the encoding of actual hypothesis, and $-\log \Pr(E | \mathbf{H}, \theta)$ measures of the classified examples given the hypothesis. This code-length can be turn into two components. Making an assumption that the hypothesis and additional parameters are *a priori* independent, $\Pr(\mathbf{H}, \theta) = \Pr(\mathbf{H}) \cdot \Pr(\theta)$, we can separate from Equation (6) all parts that depends only on θ . These parts can be effectively ignored, as it is the code-length for the unclassified examples. The rest, giving the evidence E as a sequence of N examples, \vec{e} , together with their classes \vec{c} , is outlined as [7, p58]:

$$-\log \Pr(\mathbf{H}) - \log \Pr(\vec{c} | \vec{e}, \mathbf{H}) + \frac{|\mathbf{H}|}{2} \log N + \frac{1}{2} \log I(\mathbf{H}, \theta), \quad (7)$$

where $I(\mathbf{H}, \theta)$ could appropriately be called conditional Fisher information given evidence of a single classified example.

Equation (7) can be described as that it consists of the first component that that encodes the model (hypothesis), and the second part that encodes the data using the model. A more complex hypothesis fits the data better than a simpler model, in general. We see that MML encoding gives a trade-off between hypothesis complexity, and the goodness of fit to the data.

MML naturally and precisely trades model complexity for goodness of fit. A more complicated model takes longer to state (longer first part) but probably fits the data better

(shorter second part). So an MML metric won't choose a complicated model unless that model pays for itself.

2.1 Minimum description length (MDL)

Rissanen adopts minimum encoding as a fundamental principle, and proposed *Minimum description length (MDL)* principle [7].

MDL can be explained in the context where the hypothesis space is a sequence of model classes M_1, M_2, \dots , where the model class M_k is parameterized by k real values λ . M_k thus corresponds to an hypothesis \mathbf{H} together with additional parameters θ . Given an i.i.d. evidence E , and likelihood function $f(E | M_k, \lambda)$, a method of selecting a hypothesis needs to choose both a k to select a model class and a λ to select the parameters for that model class.

One chooses \hat{k} and $\hat{\lambda}$ to minimize [7, p59]:

$$-\log f(E | M_k, \lambda) + \frac{k}{2} \log N + \left(\frac{k}{2} + 1\right) \log(k + 2) \quad (8)$$

This can be interpreted as a modified maximum likelihood approach. The second two terms modify the likelihood according to the dimensionality of the class M_k and the size N of the evidence. Because in this case $k = |\mathbf{H}| + |\theta|$, this approach is almost equivalent to choosing \mathbf{H} to minimize

$$-\log f(\vec{c} | \vec{e}, \mathbf{H}) + \frac{|\mathbf{H}|}{2} \log N + \left(\frac{|\mathbf{H}|}{2} + 1\right) \log(|\mathbf{H}| + 2), \quad (9)$$

This Equation (9) can be compared with the Equation (7) from the minimum message length framework. The MDL formula differs primarily in that it has no prior term, but also final minor quantisation term differs. Both these differing terms become insignificant for large N [6].

Both MML and MDL represent information-theoretic approach to learning. The difference between MDL and MML is a source of ongoing confusion among academics and encyclopaedia writers alike. Superficially, the methods appear mostly equivalent, but there are some significant differences, especially in interpretation:

- MML is a fully subjective Bayesian approach: it starts from the idea that one represents one's beliefs about the data generating process in the form of a prior distribution.
- MDL avoids assumptions about the data generating process altogether.

Both methods make use of two part codes: the first part always represents the information that one is trying to learn such as the index of a model class (model selection), or parameter values (parameter estimation). The second part is an encoding of the data given the information in the first part. The difference is that in the MDL literature, it is advocated that parameters that we do not want to learn

should be moved to the second part of the code, where they can be represented together with the data by using a so-called one-part code. This is often more efficient than a two-part code. In the original description of MML, all parameters are encoded in the first part so all parameters are learned.

3. EXPERIMENTS

3.1 Data

The speech database used ASR experiments is the Oldenburg Logatome Corpus (OLLO). It contains 150 different non-sense utterances (logatomes) spoken by 40 German and 10 French speakers. Each logatome consists of a combination of consonant-vowel-consonant (CVC) or vowel-consonant-vowel (VCV) with the outer phonemes being identical. To provide an insight into the influence of speech intrinsic variabilities on speech recognition, OLLO covers several variabilities such as speaking rate and effort, dialect, accent and speaking style (statement and question). The OLLO database is freely available at <http://sirius.physik.uni-oldenburg.de>.

ASR experiments were carried out with a Hidden Markov Model (HMM) with three states and eight Gaussian mixtures per HMM state. The system was set up to closely resemble, i.e. confusions could only occur for the middle phonemes. This was achieved by training and testing several HMM systems with each corresponding to a different outer phoneme. Additional delta and acceleration features were added to the 13 cepstral coefficients, yielding a 39-dimensional feature vector per time step. The ASR test set contained the same utterances as used in HSR experiments (ten speakers with 150 utterances each). Speech files from the remaining 40 speakers in OLLO were chosen for the training process. The frequency of phonemes and gender were equally distributed in the training and test set. A speech-shaped noise was added at SNR ranging from 25dB to -10dB in 5dB steps to training and test data, resulting in a matched training-test-condition.

3.2 Failure Diagnosis Procedure

For the learning process (c.f Section 2) a single list was constructed for each noise setup from ASR data. Each line (an example) consisted of a class value C and a feature vector \vec{e} . The feature vector \vec{e} consisted of:

1. Middle phoneme spoken
2. Outer phoneme spoken
3. Middle phoneme recognized
4. Outer phoneme recognized
5. String coding the speaker's identity of the belonging utterance
6. String coding the dialect for 'no dialect', 'East Frisian', 'Bavarian', 'East Phalian', and 'French' dialects

The class value $C=1$ if middle phoneme spoken and middle phoneme recognized was the same; otherwise, it was set to zero.

Note that the outer phoneme could not be confused due to the closed test setup with a given outer phoneme. The

inclusion of outer phonemes as feature component enables an analysis of coarticulation effects. The collected data was used for the learning process. The IND [8] program was used for learning and testing the failure diagnosis.

3. RESULTS

We used similar analysis with similar features in our previous work on analysis of human and automatic speech recognition [1]. Here we were specifically focused on ASR data, and in addition with speech-shaped noise. This analysis confirmed our previous findings.

Regarding dialect, the analysis of the data showed no large impact on ASR failures. The dominant feature for trees' splitting was the middle consonant of the VCV utterances. Furthermore, neither the speaker index nor the outer vowels were selected as splitting criteria which means that speaker-specific information and coarticulation effects seems to play a subordinate role for this consonant classification task.

Fig. 1 shows typical trees as produced by learning procedures. The results match quite well confusion matrices that are produced by classical ASR evaluation. Additional features outlined in Section 3.2 were not significant for the learning procedure, as they don't have an impact on classification of ASR failures. Specification of new features belongs to our future work.

PhonemeSpok = k: PhonemeRec = d: 0.08333 0.9167 0 PhonemeRec = t: 0.0625 0.9375 0 PhonemeRec = g: 0.0625 0.9375 0 PhonemeRec = k: 0.9961 0.003906 1 (a)
PhonemeSpok = k: PhonemeRec = d: 0.125 0.875 0 PhonemeRec = t: 0.0625 0.9375 0 PhonemeRec = g: 0.05 0.95 0 PhonemeRec = k: 0.9959 0.004065 1 (b)
PhonemeSpok = k: PhonemeRec = d: 0.9545 0.04545 0 PhonemeRec = t: 0.875 0.125 0 PhonemeRec = g: 0.9667 0.03333 0 PhonemeRec = k: 0.006024 0.994 1 (c)
PhonemeRec = k: PhonemeSpok = d: 0.9706 0.02941 0 PhonemeSpok = t: 0.9706 0.02941 0 PhonemeSpok = g: 0.9737 0.02632 0 PhonemeSpok = k: 0.02 0.98 1 (d)

Fig 1. Limited cut-out parts (for illustration purposes) of MML trees of phoneme "k" for SNR (a) 25dB, (b) 10dB, (c) 0dB and (d) -10dB. The last values represent the class value C .

4. CONCLUSION

We contributed to the topic of failure diagnosis of computer speech recognition, using minimum encoding style decision trees. We found that the middle consonant of the VCV utterances is the most significant factor partitioning the data on the basis of logatom recognition rate.

ACKNOWLEDGMENT

This work was supported by the of the Ministry of Education of the Slovak Republic, Scientific Grant Agency project number 2/0138/08 Applied Research project number AV 4/0006/07 and by the European Education, Audiovisual and Culture Executive Agency LLP project EURONOUNCE (URL: <http://www.euronounce.net>).



Education and Culture DG

Lifelong Learning Programme

This project has been funded with support from the European Commission. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

REFERENCES

- [1] L. Breiman, J. Friedman, Ch J. Stone, and R. A. Olshen, *Classification and Regression Trees*, Chapman & Hall/CRC, New York, 1983.
- [2] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, USA, 1993.
- [3] S. D. Peters, P. Stubble, and J. M. Valin, "On the limits of speech recognition in noise," in ICASSP '99, March 1999, vol. 1, pp. 365–368.
- [4] B. T. Meyer and M. Wachter, "Phoneme confusions in human and automatic speech recognition", *Interspeech 2007 - Eurospeech*, Antwerp, Belgium, 2007.
- [5] C. S. Wallace, *Statistical and Inductive Inference by Minimum Message Length*, Springer-Verlag, Information Sci. & Stats., May 2005.
- [6] W. L. Buntine, *A Theory of Learning Classification Rules*, Ph.D. thesis, University of Technology, Sydney, 1991.
- [7] J. Rissanen, "Modeling By Shortest Data Description", *Automatica*, 1978, Vol. 14, pp 465-471
- [8] W. Buntine, "Tree classification software," in *Technology 2002: The Third National Technology Conference and Exposition*, Baltimore, USA, December 1992.