

Diagnostics of Speech Recognition: On Evaluating Feature Set Performance

Milos Cernak^{1,2}, Mohamed Benzeghiba² and Christian Wellekens²

¹Slovak Academy of Sciences, Bratislava, Slovakia

²Institut Eurecom, Sophia-Antipolis, France

Abstract

In this paper we present an explorative study of diagnostics of speech recognition for finding subsets of features that are most informative in terms of incorrect speech recognition, if variable speech is recognized. The impact on both MFCC and PLP features is investigated. Standard HMM-GMM phoneme-based ASR system with no grammar is used for collection of the all the correct and wrong decodings, and decision tree analysis is used with questions about variance of feature coefficients in the tree nodes. The paper presents various results on importance of quefrency regions in terms of intrinsic speech variabilities, and contributes to better understanding of efficiency of used front-end.

1. Introduction

Speech recognition system depends tremendously on used feature set. Short-term features are calculated on the frame-by-frame basis directly from the acoustic waveforms, and later are used by application specific, usually HMM, engine. Selection of proper feature set belongs to the most important tasks in a design of a speech recognition system. There is an extensive literature on acoustic features for ASR and their selection (see e.g. [1]), which is still difficult task. The aim of this work is to get better understanding of the performance of the different feature sets in the terms of speech variabilities.

In speech recognition, speech variability is one of the major error sources. Speech variabilities may be classified to the two main categories: extrinsic variabilities are due to the environment (noise, telecommunication channels), and intrinsic variabilities that convey information about the speaker himself (gender, age, social and regional origin, health and emotional state) [2]. There is also a well studied impact of stressed speech on speech and speaker recognition. Stress in this context refers to speech produced under cognitive, physical, emotional stress, and stress due to presence of noise (known as the Lombard effect). Research on impact of intrinsic speech variabilities and stressed speech on speech recognition is overlapped. We have recently found a link between intrinsic speech variations and emotional speech (as a kind of stressed speech) [3].

Within the European DIVINES project (divines-project.org) we study speech recognition deficiencies in dealing with speech recognition variabilities. The ultimate goal would be to achieve better understanding of source of errors, or a signal modeling framework and robust features which are immune to the intrinsic speech variations. To achieve this goal we designed a diagnosis tool, which was already successfully applied for diagnostic purposes on the acoustic-phonetic level [4]. In this paper, we are focused on an analysis of standard feature sets (MFCC and PLP) of ASR systems, exploring impact of intrinsic speech variabilities on speech recognition.

The paper is structured as follows. Section 2 describes used diagnostic tool. Next Section 3 introduces used database and ASR systems. Section 4 overviews the process of decision tree analysis and sections 5 and 6 discuss the results and outlines future work.

2. DASR TOOL

We designed and implemented the DASR (Diagnostics of ASR) tool. Our speech recognition decoder (see Section 3 for more details) generates either ERA_IN files or CTM files¹. This gives users of the DASR Tool an opportunity to use also Lin Chase's CMU Error Region Analysis (ERA) tool, and scoring the output of speech recognizers via the NIST *sclite()* program. We found interesting to use both programs during our work on speech recognition diagnostics.

The overview of our tool is depicted in Fig. 1. The tool is implemented in MATLAB environment. In the first stage the output files of the decoder are converted and stored in an internal format, which stores all hypothesized (henceforth **HYP**) phonemes and all reference (henceforth **REF**) phonemes. Then, the initial list is split in two parts, the first containing REF sequences and the second part of HYP sequences, which are aligned using maximal substring matching. In next stage the training and testing files for decision tree analysis are generated. According to them the features of phonemes (at the acoustic, phonetic, phonologic level) are loaded or calculated, and decision tree analysis is performed. The primary technique for the analysis that the tool supports, is the CART technique described in [7]. In addition, C4.5 technique [8] is supported, as its importance for diagnostic purposes has been already shown in [9]. Trained trees may be printed in a text tabular fashion, or may be graphically displayed.

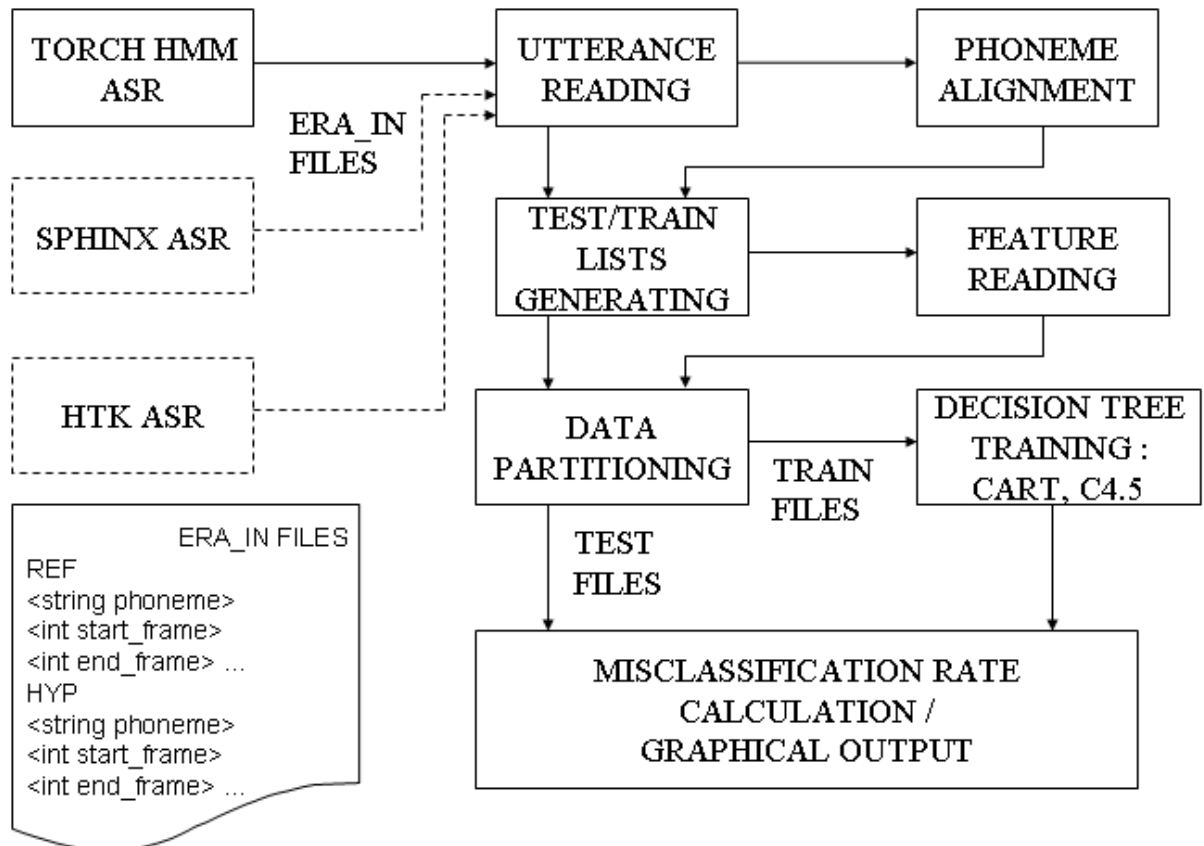


Fig. 1. DASR Tool: Diagnostics of ASR Tool. The picture highlights the main functional modules. The input is provided by speech decoder, and the tool further processes the data using decision tree analysis. Finally it provides calculation of misclassification rates on test data and depiction of trained trees. Dashed boxes are optional, showing that DASR tool should be independent of used ASR system.

¹ Both ERA_IN files (see [5] for the file definition format) and CTM files (see [6] for the file definition format).

To help categorize the errors, we use similar concepts as in [5] and [10]. Let t_{w_i} denote the start frame of the i -th phone in a transcription, then the central position of the i -th phone can be written as:

$$c_{w_i} = (t_{w_{i+1}} - t_{w_i}) / 2.$$

Using maximum substring matching algorithm we assign to each HYP phoneme one of the following categories: match, substitution, insertion or deletion. Using this information we add a label about correct or incorrect decoding as well. In addition, using two aligned sequences $\hat{\mathbf{w}}$ for decoded sequence and \mathbf{w} for reference sequence, we define w_j as the REF phoneme to the HYP phoneme w_i in the following way:

1. If w_i has a label match or substitution, we define w_j as its REF phoneme if $j = i$. Here j is an index to the REF sequence and i is an index to the HYP sequence.
2. If w_i has a label insertion or deletion, we define w_j as its REF phoneme if:

$$t_{w_j} < c(\hat{w}_i) \leq t_{w_{j+1}},$$
 where j is an index to the REF sequence, i is an index to the HYP sequence, t_{w_j} is the start frame of the REF phoneme w_j , and $c(\hat{w}_i)$ is the central position of the HYP phoneme \hat{w}_i .

In the future we are considering making this DASR tool publicly available.

3. USED DATABASE AND ASR SYSTEM

We use the OLLO database [11], which has been recorded for the purpose of study of speech recognition deficiencies in dealing with speech intrinsic variabilities. The database is designed for recognition of individual phonemes that are embedded in logatomes, specifically, CVC and VCV sequences. Several intrinsic variabilities in speech are represented in OLLO, by recording from 40 speakers from four German dialect regions, and by covering three speaker-dependent variabilities: gender, age and dialect, and six speaker-independent variabilities: fast, slow, loud, quiet, question and statement speaking styles. We used NO-accent training and testing parts of OLLO database.

Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) based speech recognition system is trained using public domain machine-learning library TORCH [6] on the training set that consists of 13446 logatome utterances. Three states left-right HMM models were trained for each of the 26 phonemes in the OLLO database including silence as well. Gaussian mixture models with 17 Gaussians per state and diagonal covariance matrices were used to model the emission probability densities of the 39 dimensional feature vectors - 13 cepstral coefficients and their derivatives (Δs) and double derivatives ($\Delta \Delta s$). The phoneme HMMs are connected with no skip. We extended the TORCH library in a package of calculation and storage of feature data, necessary for further statistic processing. The decoder collects the feature data by running on the testing set that consists of 13466 logatome utterances. We trained and tested two ASR systems, one with MFCC feature set and the second with the PLP front-end. All the features were calculated using HTK *hcopy* tool. We calculated MFCC vectors every 10 msec using windows of size 25 msec. The same settings were applied also for calculation of PLP vectors; we only used power rather than the magnitude of the Fourier transform in the binning process. Average phoneme recognition performance of the ASR systems on this task was 76.06 % (the lowest accuracy had recognition of fast speech: 71.94 %, and the highest accuracy had speech with statement style: 80.48 %). The MFCC features performed slightly better than PLP features (all our experiments were done on clean speech).

During the decoding process, both correct and incorrect decodings (cases in the terminology of decision tree analysis) are collected. The REF sequence is acquired by Viterbi forced alignment. At the end of the Viterbi computation for the last frame of the utterance the aligner stores the phone assignments to the frames, along with the actual scores associated with each segmentation.

4. DECISION TREE ANALYSIS

Decision tree analysis is performed based on the observation vectors of the MFCC and PLP coefficients (c_0, c_{1-12}), their derivatives and double derivatives). Motivated by [12], we calculated variance of the feature vectors for each HYP phoneme. Fig. 2 overviews the calculation of the 39-D phoneme feature representation used for the further analysis.

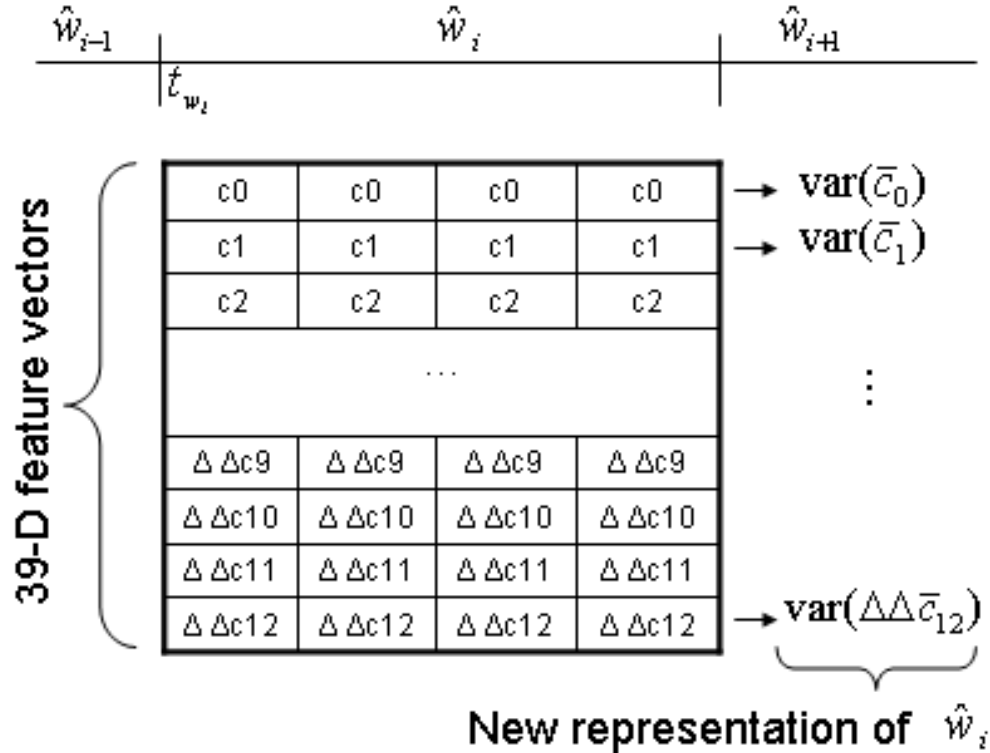


Fig. 2. Calculation of 39-D phone feature representation used in decision tree analysis. The picture shows an example of the calculation for four frames of the HYP phoneme \hat{w}_i .

Variance of speech features is calculated for each of HYP phonemes. This new 39-D parameterization is stored in the list (one item for one HYP phoneme), together with the labels about correct or incorrect decoding. These labels are later predictees for decision tree training process. We used CART technique to create six decision trees, one for each speech variability (5 variabilities plus 1 normal, statement style, speech). All the presented results in this paper were got using stopping grow criterions of minimal 10 of the cases in a terminal node and minimal entropy gain of 3%. More detailed description of the training process can be found in [4]. Splitting of the correct/incorrect cases during the training was done using questions about variances of features.

5. RESULTS

In this section, we present major results obtained from this study. We investigated both MFCC and PLP features. We went over the trained trees, following paths leading to the most probable classification of incorrect decodings. We collected all the features associated with these paths. We

can interpret these features as most significant features for prediction of incorrect decoding. The results for MFCC and PLP front-ends are shown in the Tab. 1. Decision trees for normal speech (trained on both MFCC and PLP features) have the lowest misclassification rates (the highest estimated accuracies of trained classifiers). This implies that building classifiers/predictors for correct/incorrect recognition for variable speech is more difficult. In addition, PLP decision trees have higher misclassification rates than MFCC trees. We observed that it follows the trend of lower ASR performance if PLP features are used (in clean speech).

Tab. 1. Major MFCC and PLP coefficients selected

Variability	MFCC		PLP	
	Misclassification rate [%]	Features	Misclassification rate [%]	Features
Fast	16.39	c_{12}	18.44	c_{12}
Slow	24.53	$c_{12}, c_0, c_5, \Delta\Delta c_8, \Delta c_3$	26.49	$c_{12}, c_0, \Delta\Delta(c_{12}, c_0), \Delta c_{12}$
Loud	18.26	c_{12}	22.37	$c_{12}, c_0, c_7, \Delta\Delta c_4, \Delta c_{11}$
Quiet	27.80	c_{12}	31.82	c_{12}, c_0
Question	27.43	$c_{12}, c_8, c_9, \Delta c_9, \Delta\Delta c_{10}$	26.37	$c_{12}, c_0, \Delta\Delta c_{12}, \Delta c_6, c_6$
Normal	15.27	c_{12}	17.85	c_{12}

In [13, 14] the authors shows that the lower quefrency coefficients generally have higher F-ratio (a measure of separability between multiple speech classes) and should therefore offer better class separation and so better ASR performance. Arslan and Hansen [15] have also confirmed that coefficients in the middle of quefrency region are the most relevant for dialect classification. Our findings imply that upper quefrency region (plus deltas and double deltas) is the most informative for predicting incorrect speech recognition. The most informative coefficient across all the variable speech recognition for this prediction was in our study c_{12} coefficient. For slow and questioning styled speech also dynamic features were found most informative. Dynamic features were found relevant also for loud speech in using PLP front-end.

6. IMPLICATIONS FOR SPEECH RECOGNITION

The general conclusion of this study is that the upper quefrency region and less middle region are the most informative for predicting incorrect speech recognition. Discarding the higher cepstral coefficients is sometimes normal practice in ASR. We confirmed that these coefficients are problematic. In addition, we proposed the diagnostic technique for exact specification of problematic coefficients. Some previous works confirmed different contribution of quefrency regions to recognition of stressed speech [15, 16]. New frequency scales have been there proposed, which are less sensitive to variations caused by stress without degrading the performance of neutral speech recognition. Having results of our study we confirm that upper quefrency region is also very important in terms of incorrect speech recognition.

In the future we would like to perform more experiments with extended feature set (beyond c_{12}), if our current findings will be confirmed. Moreover we would like to include these findings into feature set design, which is less sensitive to intrinsic speech variations.

Acknowledgments

This work has been supported by the EU 6th Framework project DIVINES under the contract number IST-2002-002034. The views expressed here are those of the authors only. The Community is not liable for any use that may be made of the information contained therein.

References

1. E. L. Bocchieri and J. G. Wilpon. "Discriminative feature selection for speech recognition," *ComputerSpeech and Language*, vol. 7, no. 3, pp. 229–246, July 1993.
2. M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens. "Impact of variabilities on speech recognition," in *SPECOM'2006*, Saint-Petersburg, Russia, June 2006, pp. 3–16.
3. M. Cernak and C. Wellekens. "Emotional aspects of intrinsic speech variabilities in automatic speech recognition," in *SPECOM'2006*, Saint-Petersburg, Russia, June 2006, pp. 405–408.
4. M. Cernak and C. Wellekens. "Diagnostics of speech recognition using classification phoneme diagnostic trees," in *CI 2006 (Special Session on NLP)*, San Francisco, CA, USA, November 2006.
5. Lin Chase. Error-Responsive Feedback Mechanisms for Speech Recognizers, Ph.D. thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA, April 1997.
6. R. Collobert, S. Bengio, and J. Marithoz. "Torch: a modular machine learning software library," Tech. Rep. IDIAP-RR 02-46, IDIAP, 2002.
7. L. Breiman, J. Friedman, Ch J. Stone, and R. A. Olshen. *Classification and Regression Trees*, Chapman & Hall/CRC, New York, 1983.
8. J. R. Quinlan. *C4.5: Programs for Machine Learning*, Morgan&Kaufmann Publishers, SanMeteo, CA, USA, 1993.
9. W. Buntine. "Tree classification software," in *Technology 2002: The Third National Technology Conference and Exposition*, Baltimore, USA, December 1992.
10. L. Zhang and S. Renals. "Phone recognition analysis for trajectory hmm," in *Interspeech2006 ICSLP*, September 2006.
11. T. Wesker, M. Meyer, K. Wagener, J. Anemuller, A. Mertins, and B. Kollmeier. "Oldenburg logatome speech corpus (ollo) for speech recognition experiments with humans and machines," in *Interspeech 2005*, September 2005, pp. 1273–1276.
12. D. X. Sun and L. Deng. "Analysis of acoustic-phonetic variations in fluent speech using TIMIT," in *International Conference on Acoustics, Speech, and Signal Processing, ICASSP-95.*, Detroit, USA, 1995, vol. 1, pp. 201–204 vol.1.
13. K. K. Paliwal. "Dimensionality reduction of the enhanced feature set for the hmm-based speech recognizer," *Digital Signal Processing*, vol. 2, no. 3, pp. 157–173, July 1992.
14. S. Nicholson, B. Milner, and S. Cox. "Evaluating feature set performance using f-ratio and j-measures," in *EUROSPEECH 97*, Rhodes, Greece, September 1997, pp. 413–416.
15. Levent M. Arslan and John H. L. Hansen, "A study of temporal features and frequency characteristics in american english foreign accent," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 28–40, 1997.
16. S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 4, pp. 429–442, 2000.