A PHONETIC SIMILARITY BASED NOISY CHANNEL APPROACH TO ASR HYPOTHESIS RE-RANKING AND ERROR DETECTION

Martin Hacker

Embedded Systems Initiative (ESI)
University of Erlangen-Nuremberg, Germany

Elmar Nöth

Pattern Recognition Lab
University of Erlangen-Nuremberg, Germany

ABSTRACT

We present a new method to augment the correct transcript from automatic speech recognition (ASR) output containing multiple hypotheses. The error-prone ASR process is taken as black box and modeled as a noisy channel on phoneme level. The probabilities of the individual phoneme errors are assigned according to phonetic confusability. We score potential candidate hypotheses by their posterior probability of being the channel input given the competing ASR hypotheses as observed output. The resulting scores provide useful information not included in traditional confidence measures.

We investigated the usefulness of the method for rescoring, re-ranking and word error detection. The method alone is not powerful enough to improve the recognition results, but by employing a decision tree classifier it is possible to isolate cases where the method works very well. Our results show that the combination with other knowledge sources and post-processing techniques can lead to promising improvements.

Index Terms— Automatic speech recognition, error modeling, confidence, re-ranking, error detection

1. INTRODUCTION

As it is not possible for automatic speech recognizers (ASR) to completely disambiguate the input solely with the knowledge that is accessible during the decoding phase, most systems provide alternative hypotheses in the form of n-best lists or word lattices. To increase the reliability, it is beneficial to apply post-processing techniques that can pick the correct hypothesis out of the alternatives provided by the decoder or to detect erroneous parts when none of the alternatives is correct.

Numerous approaches have been proposed to rescore [1] or re-rank recognition hypotheses both on utterance [2] and word level [3] and to detect word errors within recognition results [4, 5]. Most of these techniques use numeric and nominal features such as ASR confidence values and POS tags to train a classifier or to calculate joint confidence values that integrate different kinds of knowledge.

Nevertheless, humans show better performance in detecting ASR errors. Skantze [5] observed that human subjects benefit from information contained in n-best lists. It though remained unclear how this information can be formalized.

It seems likely that, to some extent, humans employ their experience about similar-sounding and easily confusable words to reconstruct the original utterance from the sound that is roughly reflected by the words in an n-best list (cf. also [6]). We tried to reproduce this assumed process with a formal model that integrates knowledge about phonetically similar sounds and specific error characteristics of the speech recognizer as well as phonetic information distributed over multiple competing hypotheses in n-best lists or word lattices.

2. THE METHOD

The method uses information about phonetic similarity of competing ASR hypotheses (as can be found in n-best lists, word lattices or in systems that employ parallel recognizers) and knowledge about acoustic confusability to score potential hypotheses according to their posterior probability of being the correct transcript. Instead of using the posterior probability computed by the speech recognizer we use the posterior probability of the error model proposed in the following.

We assume that the ASR process randomly transforms the reference transcript $\hat{r}$ of the input speech signal to a list of competing hypotheses $H = h_1 \dots h_n$ with probability $P(h_1 \dots h_n|\hat{r})$. Bayes' rule gives the posterior probability of a potential reference transcript r given the observed hypotheses:

$$P(r|h_1 \dots h_n) = \frac{P(h_1 \dots h_n|r)P(r)}{P(h_1 \dots h_n)}$$

For simplifying reasons, we assume that the alternative hypotheses are independent of each other (Naive Bayes assumption), i.e. we imagine that the probabilistic ASR process had been run n times with the same input. The posterior probability factorizes into a language model score $L_H(r)$ and a confusability score $C_H(r)$:

$$P(r|H) = \frac{P(r)}{P(h_1)\cdot\ldots\cdot P(h_n)} \cdot \prod_{i=1}^{n} P(h_i|r) = L_H(r) \cdot C_H(r) \quad (1)$$

$L_H(r)$ can be estimated with a language model. The probabilities $P(h_i|r)$ denote how likely the recognizer understands h_i whenever r is spoken. Abstracting from the actual acoustic realization of the speech input, they represent knowledge about general confusability. The notion of confusability can be adapted to the characteristics of a specific speech recognizer, as described in the following subsections.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2.1. The Noisy Channel Model

The noisy channel model is a well-known framework used for error correction, particularly spelling correction [7]. It comprises the imagination that the input is transmitted through a noisy channel with randomly occurring independent manipulations of items on a certain sub-level. In the following, we define phonemes to be the basic units that are manipulated independently (though other segmental units like syllables or words could be used as well). The set of potential manipulating operations involved in ASR errors is limited to substitutions, insertions and deletions[1]. We further assume that every input and output phoneme can be affected by only one operation. For clarity reasons, we model the preserving of a phoneme as substitution with itself. Thus we determine that every input/output phoneme undergoes exactly one operation.

Given a phonetic alphabet $\mathcal{P}$ and two phoneme strings $\mathcal{R} = \mathcal{R}_1 ... \mathcal{R}_l$ and $\mathcal{H} = \mathcal{H}_1 ... \mathcal{H}_m$ representing the standard pronunciation of r and h, the probabilities in $C_H(r)$ factorize as

$$P(h|r) = P(\mathcal{H}|\mathcal{R}) = \sum_{e \in E_{\mathcal{R} \to \mathcal{H}}} P(e|\mathcal{R}) \qquad (2)$$

with e iterating over all operation sequences that transfer $\mathcal{R}$ into $\mathcal{H}$ using substitutions, deletions and insertions. This is compliant with the common notion of the edit distance problem which can be efficiently solved by dynamic programming [8]. This algorithm can be easily adopted to calculate the sum in equation 2 by using the following definitions:

$$\begin{aligned} P(\mathcal{H}|\mathcal{R}) &= \alpha(l,m) \cdot (1 - p_{ins}) \\ \alpha(0,0) &= 1 \\ \alpha(i,j) &= \alpha(i-1,j-1) \cdot P(sub(\mathcal{R}_i, \mathcal{H}_j)) \\ &\quad + \alpha(i-1,j) \cdot P(del(\mathcal{R}_i)) \\ &\quad + \alpha(i,j-1) \cdot P(ins(\mathcal{H}_j)) \end{aligned} \qquad (3)$$

A variant of the method does not calculate the overall probability of all possible operation sequences but only regards the most probable operation sequence and uses its probability instead. This can be achieved by the slight modification that the sum in equation 2 and 3 is replaced by a maximum function. This variant is analogue to the Viterbi algorithm for HMMs which is commonly used by speech decoders to approximate the overall probability as both probabilities have been proven to be highly correlated for speech. We thus expect that such a correlation is also given for the confusability of hypotheses.

The probabilities of the individual phoneme errors in equation 3 can be provided by a phoneme error model and can be made dependent on the current or preceding phonemes.

2.2. Phoneme Error Models

The probabilities can be estimated from data consisting of pairs (r, h) with a reference transcript r and a hypothesis h. The first step is to make a phoneme alignment between the phonetical representations of r and h that indicates which phonemes are kept, which are inserted or deleted and which are replaced by which one. This is well known as the above mentioned edit distance problem [8]. The objective function is defined by cost weights for the individual edit operations. We explored two methods to determine the initial alignment:

UNIFORM: The weights are set to 4 for any non-identical substitution and to 3 for deletions and insertions. This is a common configuration for speech alignments (cf. [9]).

FEATURE: The phonemes are modeled by a set of individual phonological features. The weights for the edit operations are defined as a function of the number of features that are affected by the operation. This approach has been applied successfully to speech alignment before (see e.g. [10]).

In a second step, the data can be used to estimate the probabilities of interest by the corresponding counts[2] as follows[3]:

$$\begin{aligned} P(sub(\mathcal{R}_i, \mathcal{H}_j)) &= (1 - p_{ins}) \frac{\#(\mathcal{R}_i, \mathcal{H}_j)}{\#(\mathcal{R}_i, *)} \\ P(del(\mathcal{R}_i)) &= (1 - p_{ins}) \frac{\#(\mathcal{R}_i, \epsilon)}{\#(\mathcal{R}_i, *)} \\ P(ins(\mathcal{H}_j)) &= p_{ins} \frac{\#(\epsilon, \mathcal{H}_j)}{\#(\epsilon, *)} \\ p_{ins} &= \frac{\#(\epsilon, *)}{\#(*, *)} \end{aligned}$$

The resulting probabilistic phoneme error models supply the probabilities of the noisy channel for the algorithm proposed in 2.1 (Eq. 3). By using the Viterbi-like variant of the algorithm, it is possible to refine the initial alignment and hence the error model. By iterative application of these refinement steps, following the paradigm of the EM-algorithm, we generated two further models UNIFORM-EM and FEATURE-EM.

2.3. Normalization

The scores defined in equation 1 heavily depend on the number n of competing hypotheses and the length of the strings. To make the values comparable between different utterances they are normalized as follows:

$$\overline{L}_H(r) = \frac{P(r)^{1/len(r)}}{(P(h_1)^{1/len(h_1)} \cdot .... \cdot P(h_n)^{1/len(h_n)})^{1/n}}$$

$$\overline{C}_H(r) = \prod_{i=1}^{n} P(h_i|r)^{1/(k_i n)} \quad ; \quad \overline{S}_H(r) = \overline{L}_H(r) \cdot \overline{C}_H(r)$$

with k_i being the number of edit operations for the most probable operation sequence to transduce r into h_i.

3. RELATED WORK

Ristad [11] presented a stochastic model for the adaptation of string-edit distance to data. It is compatible with UNIFORM-EM (see 2.2) and applies the same EM-estimation strategy.

Kemp/Schaaf [12] and Mangu/Brill [13] proposed methods for calculating confidence scores from word lattices. Both

[1] Methathesis (segment order errors) known from humans can be ruled out since the sequential frame order is preserved during the whole ASR process.

[2] It can happen that multiple alignments have the same overall cost value. In this case, we took all for the estimation step, but the corresponding counts were weighted by $1/k$ with k being the number of alternative alignments.

[3] In deviation from the description, we applied a simple smoothing technique to avoid null probabilities for unseen operations and a back-off estimation for substitutions of phonemes that were never substituted in the data.

follow the idea of hypothesis density which is implicit in our model as well. However, their approaches exploit ASR posterior probabilities, or entropy measures, respectively, while the method described here focusses on phonetic similarity.

In section 5.2, we demonstrate how our method can be applied to ASR system combination. A well-known technique for system combination is ROVER [14] which constructs word transition networks from multiple-recognizer outputs and chooses the best path by a voting procedure. Our method uses phonetic similarity as an additional information source.

The work most similar to the approach described here is the ASR post-processing technique of Ringger/Allen [15]. Like our model, it relies on a noisy channel perspective on speech recognition errors. The major differences are that Ringger's model does not account for multiple competing hypotheses at once and operates on word level. In section 5.2, we compare our method with Ringger's tool SpeechPP.

4. SPEECH CORPORA AND TOOLS

The experiments were conducted on both spontaneous casual speech (dialogue systems) and read speech (newspaper texts).

PAC and IISAH dialogue corpora (German language): The PAC corpus consists of 544 on-talk user moves collected with a Wizard-of-Oz variant of the pedestrian assistance system ROSE [16] which offers a mixed-initiative spoken language interface for tasks ranging from information retrieval (e.g. public transport live timetable questions) to more complex problem solving issues such as navigational assistance, re-planning or recommendation of locations and activities.

To evaluate if the technique presented here is adaptable to other domains, we used in addition a subset (1185 moves) of the FAU IISAH corpus [17] containing dialogues of elderly people with a speech-controlled home assistance system.

WSJ read data: The Wall Street Journal Corpus (WSJ0)[4] [18] was divided – as in the Nov92 ARPA CSR Benchmark Tests [19] – into a training set (used for LM training), a development test set (1805 moves, used for error model and decision tree training) and an evaluation test set (1285 moves).

Speech recognition and evaluation: For speech recognition we used Google's LVCSR service [20] and, for comparison, Sympalog's[5] recognition engine SymRec [21] (PAC data only) with domain-specific 7K word bigram language model.

Table 1 shows the utterance correctness rate (UCR), word (WER) and phoneme (PER) error rates of the first hypothesis as well as the 5-best oracle correctness rate (UCR-O) and the 5-best mean rank of the first correct hypothesis (MRFC).

5. RESULTS

In the following experiments we used several features computed using the model. Table 2 provides an overview. All reported results have been achieved using the normalized Viterbi variant of the scores and the FEATURE-EM model which yielded slightly better figures than other combinations.

Table 1. *ASR performance before applying the method.*

Dataset	Lang	UCR	UCR-O	WER	PER	MRFC
PAC-Google	DE	34.2%	43.4%	38.29%	23.96%	1.33
PAC-SymRec	DE	30.9%	38.8%	42.35%	31.90%	1.33
IISAH-Google	DE	48.6%	56.8%	32.55%	20.52%	1.23
WSJ-Google	EN	11.1%	16.3%	24.62%	12.01%	1.52

Table 2. *Overview: Features in the re-ranking experiments.*

NEWBEST_OLDRANK:	The original ASR rank of the best-scored ASR hypothesis.
NEWBEST_SCORE:	The score $argmax_i \overline{S}_H(h_i)$ of the best-scored ASR hypothesis.
OLDBEST_SCORE, SECOND_SCORE, WORST_SCORE:	Analogous for the first ASR hypothesis, the second-best scored and the worst-scored hypothesis.
OLDBEST_SCORERATIO, SECOND_SCORERATIO, WORST_SCORERATIO:	The ratio (relative distance) between the corresponding score and BEST_SCORE.
OLDBEST_SCOREREL, SECOND_SCOREREL:	Ratio between the absolute distances of the corresponding score and BEST_SCORE, and WORST_SCORE and BEST_SCORE.
ASR_CONFIDENCE:	The confidence of the speech recognizer.

5.1. Hypothesis rescoring

In general, rescoring of ASR hypotheses is achieved by combining confidence measures from different knowledge sources. We show the general usefulness of the above scores for this task by reporting their correlation with error measures.

The score $\overline{C}_H(h)$ shows a strong empirical linear correlation (0.879) with the phonetic similarity $P(h|\hat{r})$ of hypothesis h and reference $\hat{r}$, with the PER (-0.626) and yet a medium correlation (-0.506) with the WER of the hypothesis.

These correlations are higher than the respective correlations of the ASR confidence value (PER: -0.469 / WER: -0.453). At the same time, there is only a weak correlation between the mentioned score and the confidence (0.231). These results indicate that the score contains useful information that is different from that in the ASR confidence value.

Word-level scores: The approach can also be applied to score words in a hypothesis. To achieve this, we compute all pairwise alignments of the phonetic strings of the hypotheses in a n-best list. The alignments can be used to find the corresponding sub-strings of the competing hypotheses for a given word in one hypothesis (Fig. 1). The method described in section 2 is then applied to this set of competing phonetic sub-strings.

Similar to the utterance-level scores used above, the word-level scores show a strong correlation (0.866) with the phonetic similarity to the corresponding phonetic segment of the reference transcript.

5.2. N-Best List Re-Ranking

In the next experiment, we reordered the hypotheses in the ASR 5-best[6] list according to their score[7] $\overline{S}_H(h_i)$. Both UCR

[4]Only channel 1 and recordings with no verbal punctuation were used.

[5]http://www.sympalog.de

[6]Up to 30 n-best hypotheses were used to calculate the scores, but only the 5 best were re-ranked as the tail is unlikely to contain the correct transcript.

[7]Hypotheses with the same score were kept in the original order.

vElC@n'o: **p6 nho:f'** aUsgaN	welchen opern hof ausgang
vElC@n'U nt**ba:nho:f'** aUsgaN	welchen und **bahnhof** ausgang
vElC@nhu: **b6 ho:f'** aUsgaN	welchen huber hof ausgang

Fig. 1. *Calculation of word scores. R: n-best list with word of interest. L: phonetic alignments w/ corresponding segments.*

Table 3. *Results of the re-ranking experiments.*

	UCR	relat.	WER	relat.	MRFC
a) WSJ – Google:					
Baseline	**11.1%**		**24.62%**		**1.52**
Re-ranking	10.4%	-6.3%	24.58%	-0.2%	1.72
w/classifier	**12.1%**	+9.0%	**23.98%**	-2.6%	**1.54**
Gold standard	16.3%	+46.8%	21.05%	-14.5%	1.00
b) PAC – Google:					
Baseline	**34.2%**		**38.29%**		**1.33**
Re-ranking	18.0%	-47.4%	42.46%	+10.9%	2.27
w/classifier	**36.0%**	+5.3%	**37.81%**	-1.3%	**1.24**
Gold standard	43.4%	+26.9%	32.29 %	-15.7%	1.00
c) PAC – Combined ASR engines:					
Re-ranking	24.1%	-38.3%	35.95%	-6.1%	1.92
w/classifier	**39.5%**	+15.5%	**33.78%**	-11.8%	**1.36**
Gold standard	48.9%	+43.0%	25.65 %	-33.0%	1.00

and WER of the first hypothesis as well as the mean rank of the first correct hypothesis were considerably worse than for the original ranking (see Table 3, row "Re-ranking").

We investigated whether it is possible to augment in which cases the method improves the ranking. For this purpose, we marked all cases where the 1-best WER improved with the class label *RERANK* and all cases where it became worse with *KEEP*. We collected a number of potentially informative features (see Table 2) to train a decision tree classifier (C4.5) to distinguish the two classes. A subset of up to 4 features, including the ASR confidence, showed the best performance (Fig. 2 shows the tree for WSJ).

Using the classifier to decide when to apply the re-ranking method, UCR and WER can be improved compared to the original ranking. Table 3 a) and b) summarize the results (10-folds cross validation). It should be mentioned that most of the 5-best lists do not contain a correct hypothesis which explains the relatively low gold standard (oracle re-ranking). About 20% of the possible improvement could be achieved. A similar pattern arises when the method is applied to the SymRec 5-best lists and the IISAH data.

Combination of ASR engines:

We investigated whether the re-ranking approach can be applied to joint n-best lists from different speech recognizers (system combination). For this purpose, we merged the n-best lists of Google and SymRec according to the following procedure: take the first Google hypothesis as first, first SymRec as second, second Google as third, and so forth; prune the joint list to get equal numbers of contained hypotheses from both sources.

Table 3 c) shows the results. In only 10% of the cases where the Google result contained no correct hypothesis, the second recognizer provided one. However, the supplemental phonetic information given by the second recognizer was useful enough to considerably improve the overall performance of the classifier-based approach.

Table 4. *Combination with error correction (SpeechPP).*

	WSJ-Google: UCR	WER	*PAC-Google:* UCR	WER
ASR 1-best	**11.1%**	**24.62%**	**34.2%**	**38.29%**
Re-ranking only	12.1%	23.98%	36.0%	37.81%
SpeechPP only	11.6%	23.51%	37.7%	**35.02%**
Combined	**12.3%**	**23.16%**	**38.8%**	35.40%

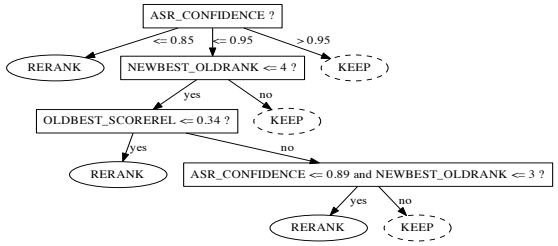

Fig. 2. *Decision tree for the application of the re-ranking.*

Combination with error correction:

We also compared our method with Ringger's error correction tool SpeechPP [15] mentioned in section 3. The two approaches tend to be complementary: The best results were achieved by a combination of both algorithms (see Table 4).

5.3. Word error detection

Word error detection can be viewed as the task to assign the labels *correct* vs. *incorrect* to each word in an ASR hypothesis. We trained a logistic regression model [22] for this problem using similar features as suggested in [4] and [5]. The classifier achieved 70.63% correctness (10-folds cross validation). Including the features from Table 2 yields a slight increase to 71.9%. The best results were achieved with a combination of OLDBEST_SCORE and OLDBEST_SCOREREL. The figures indicate that the method described in this paper can facilitate error detection, yet there is need for further research to provide a feasible technique for reliable identification of incorrectly recognized words.

6. SUMMARY

We described and evaluated our approach to augment the correct transcript from extended ASR output. The model is built upon phoneme error probabilities that can easily be estimated from data. We evaluated some applications for post-processing of ASR results from conversational dialogue systems as well as WSJ read data. It is possible to apply machine learning to the task of deciding when the method should be applied. An improved overall performance was achieved for a re-ranking strategy in combination with a decision tree classifier, particularly when multiple ASR engines were combined.

The encouraging results suggest that it might be promising to further optimize the method and combine it with other approaches to rescoring, error detection and error correction. Future investigations should evaluate the use of syllable or higher-order phoneme error models with probabilities dependent on the phoneme context or the rank of the hypothesis.

7. ACKNOWLEDGEMENTS

We would like to thank Eric Ringger for providing his source code.

8. REFERENCES

[1] Hui Jiang, "Confidence measures for speech recognition: A survey," *Speech Communication*, vol. 45, no. 4, pp. 455 – 470, 2005.

[2] Zhengyu Zhou, Jianfeng Gao, Frank K Soong, and Helen Meng, "A Comparative Study of Discriminative Methods for Reranking LVCSR N-Best Hypotheses in Domain Adaptation and Generalization," in *Proceedings of ICASSP 2006*, vol. 1, p. I.

[3] Daniele Falavigna, Roberto Gretter, and Giuseppe Riccardi, "Acoustic and word lattice based algorithms for confidence scores," in *Proceedings of ICSLP 2002*.

[4] Georg Stemmer, Stefan Steidl, Elmar Nöth, Heinrich Niemann, and Anton Batliner, "Comparison and combination of confidence measures," in *TSD '02: Proceedings of the 5th International Conference on Text, Speech and Dialogue*, 2002, pp. 181–188.

[5] Gabriel Skantze and Jens Edlund, "Early error detection on word level," in *Proc. of ITRW on Robustness Issues in Conversational Interaction*, Norvich, UK, 2004.

[6] Teresa Zollo, "A study of human dialogue strategies in the presence of speech recognition errors," in *Working Notes AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, 1999.

[7] Eric Brill and Robert C Moore, "An improved error model for noisy channel spelling correction," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL)*, 2000, pp. 286–293.

[8] Robert A. Wagner and Michael J. Fischer, "The string-to-string correction problem," *Journal of the ACM (JACM)*, vol. 21, no. 1, pp. 168–173, Jan. 1974.

[9] David S Pallet, William M Fisher, and Jonathan G Fiscus, "Tools for the analysis of benchmark speech recognition tests," in *Proceedings ICASSP*, 1990, pp. 97–100.

[10] Grzegorz Kondrak, "A new algorithm for the alignment of phonetic sequences," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, Stroudsburg, PA, USA, 2000, NAACL 2000, pp. 288–295, ACL.

[11] Eric Sven Ristad and Peter N. Yianilos, "Learning string-edit distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 5, pp. 522–532, May 1998.

[12] Thomas Kemp, Thomas Schaaf, et al., "Estimating confidence using word lattices," in *Proc. Eurospeech*. Rhodes, Greece: ESCA, 1997, vol. 2, pp. 827–830.

[13] Lidia Mangu, Eric Brill, and Andreas Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[14] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. of ASRU*, 1997, pp. 347–354.

[15] Eric K Ringger and James F Allen, "Robust error correction of continuous speech recognition," in *Proceedings of the ESCA-NATO Robust Workshop*, 1997.

[16] Bernd Ludwig, Bjørn Zenker, and Jan Schrader, "Recommendation of personalized routes with public transport connections," *Intelligent Interactive Assistance and Mobile Multimedia Computing*, pp. 97–107, 2009.

[17] Werner Spiegl, Korbinian Riedhammer, Stefan Steidl, and Elmar Nöth, "FAU IISAH corpus – a German speech database consisting of human-machine and human-human interaction acquired by close-talking and far-distance microphones," in *Proceedings of LREC*, 2010.

[18] Douglas B Paul and Janet M Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the Workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[19] David S. Pallett, Johathan G. Fiscus, William M. Fisher, and John S. Garofolo, "Benchmark tests for the DARPA spoken language program," in *Proceedings of the workshop on Human Language Technology*, Stroudsburg, PA, USA, 1993, HLT '93, pp. 7–18, Association for Computational Linguistics.

[20] Mike Schuster, "Speech recognition for mobile devices at Google," *PRICAI 2010: Trends in Artificial Intelligence*, pp. 8–10, 2010.

[21] Elmar Nöth, Axel Horndasch, Florian Gallwitz, and Jürgen Haas, "Experiences with commercial telephone-based dialogue systems," *it–Information Technology*, vol. 46, no. 6/2004, pp. 315–321, 2004.

[22] Saskia Le Cessie and JC Van Houwelingen, "Ridge estimators in logistic regression," *Applied statistics*, pp. 191–201, 1992.