DIAGNOSTICS OF SPEECH RECOGNITION USING CLASSIFICATION PHONEME DIAGNOSTIC TREES

Milos Cernak and Christian Wellekens
Department of Multimedia Communications
Institut Eurecom, 2229 Route des Crêtes - BP 193
06904 Sophia-Antipolis, France
emails: {cernak,wellekens}@eurecom.fr

ABSTRACT

More than three decades of speech recognition research resulted in a very sophisticated statistical framework. However, less attention was still devoted to diagnostics of speech recognition; most previous research report on results in terms of ever-lower WER in various intrinsic or environmental conditions.

This paper presents a diagnostics of the decoding process of ASR systems. The purpose of our diagnostics is to go beyond standard evaluation in terms of WERs and confusion matrices, and to look at the recognized output in more details. During the decoding phase, some specific data are collected at the decoder as possible causes of errors, and later are statistically analyzed using classification and regression trees. Focusing on pure acoustic phone decoding without language modeling, we present and discuss the results of the diagnostics that is used for an analysis of impact of intrinsic speech variabilities on speech recognition.

KEY WORDS

Fault diagnosis, speech recognition, intrinsic speech variabilities.

1 Introduction

Most of research in the field report results in terms of ever-lower WER acquired over some baseline, leaving questions about the causes of failures open. Evaluation of recognizer performance is usually expressed in terms of few figures like WER and confusion matrix. Diagnostics complements the evaluation[1]. In terms of ASR technology, diagnostics is the identification and more challenging, the understanding of incorrect speech recognition. Diagnostics of speech recognition should provide error patterns of the decoding process as well as of the training process. This paper aims to contribute to the diagnostics of decoding in speech recognition.

Recognition may be studied in detail considering different linguistic or phonetic properties [2]. The recognition results are usually identified using the acoustic-phonetic classes [3, 4]. Some authors go further and try to find a reason of phoneme confusion, or even their deletions and insertions. In a recent work [5], authors explored some articulatory properties of confused consonants. Comparing human and computer speech recognition, they concluded, that voicing information should actually be used for better performance of machine speech recognition. In our work we use a decision tree analysis, following work of [6, 7, 8]. The idea is to incorporate statistics of building decision trees for finding factors that cause the systematic recognition errors.

This paper aims to extend the idea of using decision trees for ASR diagnostics, and proposes phoneme diagnostic trees (PDTs). The analysis is done at the phoneme level, and provides detailed results that were not available in the previous approaches. The aim is to present a general methodology, applicable to any analysis level (such as acoustic-phonetic, linguistic) and in any variability. For each basic phoneme used in an ASR system a PDT is constructed, which links incorrect recognitions of the phoneme with a-priori specified sources of errors, hereafter called factors or phoneme features[2]. During tree building only features are selected that are statistically significant, and so main sources of errors are found. Going over the PDT, the values of the features are found, and this indicates the reasons of ASR failures.

The paper is structured as follows. Section 2 describes selection of phoneme features. Section 3 overviews the process of building PDTs, and section 4 depicts the use of PDTs. Finally section 5 discusses the contribution of the paper, and outlines future work.

2 Selecting phoneme features

Within the European DIVINES project (divines-project.org), we study speech recognition deficiencies in dealing with speech intrinsic variabilities as opposed to using prior restrictive knowledge (for instance in the form of very specific grammars). Therefore a specific speech corpus, the OLLO database, has been recorded for that purpose [9]. The database is designed for recognition

[1] Evaluation (see e.g. [1]) is an assessment of the system, measuring some parameters of the system , while diagnostics is a computing mechanism to identify faults of the system.

[2] We use the term 'phoneme features' in this paper for various representations of the phoneme; not be confused with the features extracted from the waveform such as MFCC, PLP.

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.

Speech features might be seen in three levels: the discrete phonological representation of an utterance, the acoustic pattern that results from the utterance, and the articulatory gestures that create the links between the phonological and acoustic representations. Our phonological representation is constrained to remain on the phoneme level. In the first round we have specified the following features:

- *Ph.left*: the left context, $Ph.left \in \{1, ..., L\}$, where L is number of phonemes.
- *Ph.right*: the right context, $Ph.right \in \{1, ..., L\}$, where L is number of phonemes.
- *Ph.part*: the phoneme's part, $Ph.part \in \{onset, coda, onset - coda\}$; it specifies the failures of recognition of that part.
- *Ph.duration*: the duration of the phoneme.

For the acoustic representation we have selected two measures:

- *Ph.AS_norm*: The acoustic score at each frame within the decoded phoneme, normalized by the acoustic score get from Viterbi alignment. The standard acoustic normalization scheme suggested is the Bayes' rule:

  $$P(ph|A) = \frac{P(ph) \times P(A|ph)}{P(A)}. \quad (1)$$

  But for the purposes of determining acoustic relationship between decoded and referenced recognition, we use similar measurement:

  $$Ph.AS\_norm = \frac{P(A|ph)}{P(A)_{FA}}, \quad (2)$$

  where $P(A)_{FA}$ is the acoustic score of aligned speech segment got from Viterbi alignment (see detection of incorrect recognition in Sec. 3.1). Because we have not used any grammar, $P(ph) = 1$.

- *Ph.NAQ*: The calculation of Normalized Amplitude Quotient (NAQ) [10]. This was used for voice quality representation, as it has been found to be very robust parameterization of the glottal flow. Our previous work on speech variabilities also has found this representation useful for assigning emotional aspects to each of speech variability [11]. The amplitude quotient AQ is computed as $\mathrm{AQ} = f_{AC}/d_{peak}$, where f_{AC} denotes AC flow, the part of the glottal flow that varies in time, and d_{peak} denotes negative peak amplitude of the differential flow. The AQ quantifies the closing phase of the glottal flow, and it reflects changes that occur in the glottal source when vocal intensity or phonation type is altered [10]. The above time-length measure is further normalized with respect to the length of the pitch period:

  $$\mathrm{NAQ} = \mathrm{AQ}/T, \quad (3)$$

  where T denotes the length of a pitch period.

We look at the phoneme features as at the sources of information, which could reveal possible causes of wrong recognition. From this point of view it is necessary to select such features, which are relevantly related to the possible improvements of speech recognition. Otherwise the result of the diagnostics would be useless. The above selected measures are more or less intuitive. Many researchers believe, that context and coarticulation are what make speech recognition difficult [12] (measures *Ph.left* and *Ph.right*). The measure *Ph.part* is used to reveal the affected parts of recognized phonemes, leading us to a specification of affected HMM states in trained acoustic models. The measure of phoneme duration *Ph.Dur* captures bad modeling of duration. The acoustic measure *Ph.AS_norm* is zero, if there is no difference between reference and hypothesized HMM state sequence. For erroneous regions it is positive if the acoustic score of a hypothesized sequence is less than the one of its reference sequence. The second acoustic measure *Ph.NAQ* characterizes the vocal quality, or laryngeal phonation style of the recognized phoneme, and incorporates the information from speech production as possible cause of error in speech recognition.

We have also experimented with the articulatory descriptors of the phonemes. We used Withgott and Chen's phonological descriptors [13] of dimension 25 for that purpose. Even though it was found that the Hamming distance between these phonological descriptors are powerful if it is used e.g. as phone confidence annotators [7], it cannot be used as such. During tree building, this phonological distance is predominantly selected by a training process as the most significant factor. This is due to the fact, that the distribution of that distance is exactly the same as the distribution of the incorrect decodings. We have therefore excluded this measure from further processing. We applied the same conclusion for the number of occurrences of a phoneme in the acoustic training data.

3 Building PDTs

3.1 Used ASR system

Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) based speech recognition system is trained using public domain machine-learning library TORCH [14] on the NO-accent training part of the OLLO 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 MFCC + 13 deltas + 13 accelerations). The phoneme HMMs are connected with no skip followed by decoding. 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 NO-accent testing part of the OLLO database that consists of 13466 logatome utterances. Average phoneme recognition performance of the ASR system 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 %).

Detection of wrong recognition is done by a comparison of the phonetic sequence produced by the recognizer against the expected sequence. But in addition, the time boundaries of the phonemes are taken into account, and the correctly decoded phoneme with bad duration is considered as wrong decoding. Frame tolerance of two frames is allowed for detection of that bad duration. Expected 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 frames, along with the actual scores associated with each segmentation. This acoustic score is later used as a normalization factor $P(A)_{FA}$ in Eq. 1. Having expected sequence and phoneme-aligned decoded sequence, we defined wrong recognition as described above.

Our diagnostics is based on decision trees. We split the feature data D into three groups: 50 % of the data as a training set $DATATRAIN$, 25 % of data as a test data $DATATEST$ for testing during stepwise tree building, and 25 % of the data T for testing of that trained tree. During testing of the trees we calculate the misclassification rate of the built trees on the test data T. The misclassification rates of the trees varied in the range 60-90 %.

3.2 Training

Decision trees describe how a given input can correspond to specific outputs, as a function of some factors. At each non-terminal node, there is a question requiring a binary answer about the value of the factor associated with the node. Terminal nodes, or leaves, are associated with a specific output. Such decision trees can be automatically obtained by classification and regression (CART) training technique [15], which automatically allows the most significant factor to be statistically selected using a greedy algorithm. In order to build a tree, one needs a training set of samples $DATATRAIN = \{(p_1, c_1, \vec{x_1}), ..., (p_N, c_N, \vec{x_N})\}$, where $p_n \in \{1, ..., L\}$ are phonemes that should be recognized, $c_n \in \{1, ..., L\}$ are classes (tree outputs, in our case the incorrect recognized phonemes), and $\vec{x_n} = (x_1^n, x_2^n, ..., x_M^n)$ are callculated feature vectors for that cases, each with a total of M feature measurements. Each learning sample denotes a single incorrect recognition. A training technique of CART is used to produce L decision trees, which we further call phoneme diagnostic trees (PDTs).

We build PDTs in a stepwise fashion using the `wagon` tool of the Edinburgh Speech Library with options `'-stepwise -test DATATEST -stop 4'`. In this case instead of considering all features in building the best tree, we incrementally build trees looking for which individual feature best increases the accuracy of the built tree on the provided test data $DATATEST$. Unlike within the tree building process where we are looking for the best question over all features, this technique puts a limit on which features remain available. It first builds a tree using each of the features provided, looking for which individual feature provides the best tree. The selecting that feature is builds $M - 1$ trees with the best feature from the first round with each of the remaining features. This process continues until no more features add to the accuracy or some stopping criteria is reached. We used stopping criteria of minimal 4 samples (phonemes) per terminal node. This stepwise technique is also a greedy technique but it was found that when many features are presented, especially when some are highly correlated with each other, stepwise building produces a significantly more robust tree on external test data. It also typically builds smaller trees.

During three building, the following impurity measure[3] was used. Let us define t as a node of tree. Because we build PDTs only for classification/prediction of wrong phoneme recognition (with categorical values consisted of names of incorrect recognized phomenes), the impurity $i(t)$ for the node t is calculated as:

$$i_{cat}(t) = -\sum_{c \in L} P(c|t) \times \log_2 P(c|t), \qquad (4)$$

where $p(c|t)$ is the probability of the class c in the node t.

4 Results on Using PDTs

In order to have statistically significant data, we merged all data collected by decoder for all variabilities. The training procedure described in the previous section 3.2 results in 26 decision trees (26 is the number of used phonemes in OLLO database). Each PDT has in each of its leaves again exactly 26 phonemes, with assigned probability of the specific incorrect recognition. We can observe two mutually dependent informations:

- *Selected features*: The CART training procedure selects statistically significant features, which best describe the relation of a reference phoneme and the decoded phonemes.

[3] Impurity of a set of samples is designed to capture how similar the samples are to each other. The smaller the number the less impure the sample set is.

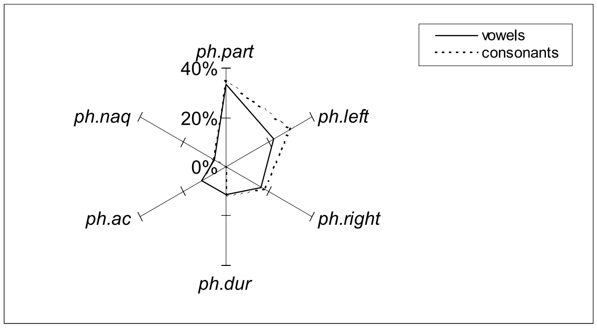

Figure 1. Relative feature rate used in PDTs, split in accordance with vowel/consonant categorization .

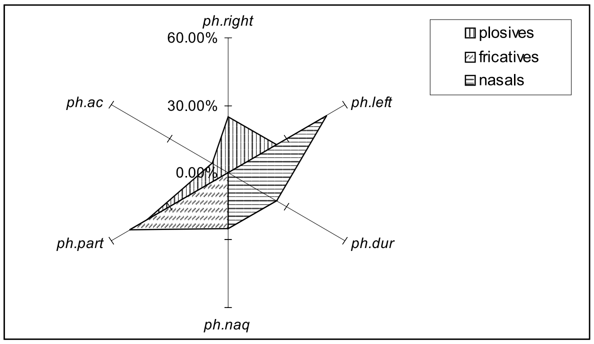

Figure 2. Relative feature rate used in PDTs of consonants, split in accordance with the manner of articulation. The part of *Ph.NAQ* and *Ph.Dur* is shared between nasals and fricatives.

- *Distribution of incorrect decodings*: Analyzing the leaves of the PDTs, one can see the most probable incorrect decodings given the selected questions about the features in the nodes leading to the analyzed leave.

As an example of possible outcome from the diagnostics, we did the following evaluation. Having 26 trained PDTs (one for each phoneme), we grouped them with respect to (a) vowels and consonants, and (b) with respect to splitting the consonants according to their manner of articulation. Going over the PDTs, we found features in questions, and calculated their relative rate of selection by the training process. Figure 1 refers to the (a) case and Figure 2 to the (b) case respectively. For both vowels and consonants, the distribution of selected features is almost identical. Only for consonants the context plays more important role than for vowels. The different picture is found when we are closely looking at consonants. The most important cause of errors for plosives is the plosive's context. For nasals the vocal quality feature *Ph.NAQ* is significant, and for fricatives, the duration of recognized phoneme.

An example of trained PDT for the phoneme /g/ is shown at Fig. 3. At each non-terminal node, there is a question requiring a yes (right) or no (left) answer about the value of the selected feature. Terminal nodes, or leaves, are associated with a specific set of probabilities for each of 26 phonemes. For the presentation purpose, only the first three most significant probabilities are depicted.

Variability	Part	Left	Right	Dur	AS	NAQ
Fast	x		x			
Slow				x	x	
Loud			x			x
Quiet		x				
Quest. style	x		x	x		
Normal style					x	

Table 1. Chief features selected for classification of ASR failures during decoding of phoneme /o/

4.1 Impact of Speech Variabilities

One of the most important aims of the DIVINES project is to understand an impact of intrinsic speech variabilities on speech recognition. Here, we show an example how PDTs can be used to this aim. Now, the training set has to be split in to 6 parts, each representing one of the six different articulation characteristics recorded in OLLO database. These characteristics include:

- speaking rate: fast, slow, normal
- speaking effort: loud, quiet, normal
- speaking style: question or statement

Due to the necessity of large amount of data for statistical training, we have selected as next example the phoneme /o/ for which the decoder collected the most data. This phoneme belongs to the worst decoded phonemes, where almost every third decoding of /o/ was incorrect because of its substitution with the phoneme /u/.

We constructed six PDTs, one for each variability, and looked at the main features selected by CART training. This information is shown in table 1.

The most significant features for fast speech is context, when the reason might be that faster speech rate may lead to more frequent and stronger pronunciation changes [2]. On the contrary, causes of errors for slow speech seem to be in bad modeling of duration and weaker discrimination of acoustic models. Loud speech shows the importance of Normalized amplitude quotient, because the speaking effort is changed here. Finally, for question style the context together with duration are the most significant, and for normal speech the only possible cause of incorrect decoding of /o/ has been found in less discriminant acoustic modeling against acoustic models of the phoneme /u/.

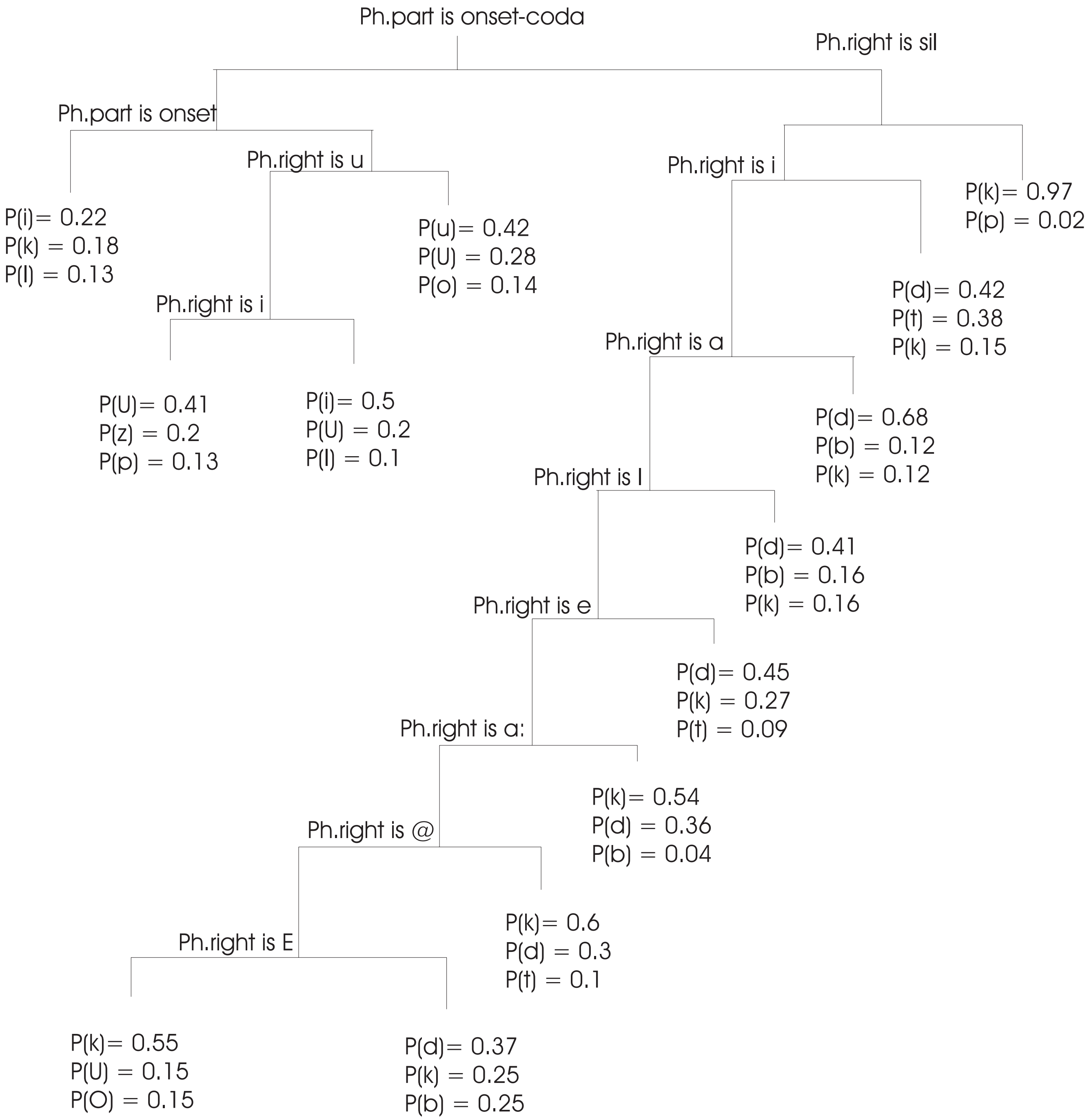

Figure 3. Decision tree of phoneme /g/. The most significant diagnostic results on the right is the misrecognition (a) of /g/ as /k/, if silence follows /g/ - the whole /g/ is misrecognized, (b) of /g/ as /d/, if /a/ follows /g/ - again the whole /g/ is affected. Looking on the left of the tree, we see an interesting impact of vowels /U/ and /I/ followed by /g/. Here the part of /g/ is decoded as /U/ or /I/ respectively. Concluding the selected main features all together, the right context of the consonant /g/ plays the most significant role in its incorrect recognition, and must be better modeled in situations, which were revealed by the diagnostic tree.

5 Conclusion

The technique of PDTs provides understanding about causes of error for individual phonemes. We believe that it is reasonable, while there are some clues in the literature, which shows that humans decode syllables as independent phone units over time [12, 16]. H. Fletcher and recently J.B. Allen defined the articulation score, a term of human speech perceptual concept, as the probability of identifying nonsense speech sounds. This definition is in fact also the definition of accuracy in machine speech recognition. We believe that using results of PDTs, using simple comparision as stated above, one can make positive contribution to the man-machine comparison. It is therefore worth to study failures of individual phonemes in details, instead of averaging recognition results over some generalized categories. Human speech recognition is for real life applications much more robust than computer speech recognition. Precise diagnostics of computer speech recognition for a given experiment may improve understanding of failures, which may further suggests modifications for next experiments, converging to the human performance.

During the training of PDTs using CART technique we have found its high sensitivity to the settings of training parameters. In our work we excluded the data sets with too few examples, and we also excluded trained trees with too high misrecognition rate achieved over the test data set. Moreover we used a stepwise building method, which produces significantly more robust trees. However, we still see room here for next improvements of the tree building process.

Motivated by [8], in the future we would like to extend current feature set. The advantage of this diagnostics is, that at the end of training we get a set of significant features, but also the rest is interesting - which features are not significant. We can simply add also more experimental features, almost everything what can be measured and has a meaning towards the reaching limits of machine speech recognizers.

6 Acknowledgments

This work has been supported by the EU 6^{th} 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.

7 References

[1] S.J. Young and L.L. Chase, Speech recognition evaluation: a review of the U.S. CSR and LVCSR programmes, *Computer Speech & Language*, 12(4), 1998, 263-279.

[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 speech variabilities on speech recognition, *Proc. of 11th Int. Conf. Speech and Computer*, Saint-Petersburg, Russia, 2006.

[3] G. Chollet, A. Astier, and M. Rossi, Evaluation the performance of speech recognitzers at the acoustic-phonetic level, *Proc. of IEEE Conf. on Acoustics, Speech, and Signal Processing*, Atlanta, USA, 1981, 758-761.

[4] M. Hunt, Speech recognition, syllabification and statistical phonetics, *Proc. of ICSLP*, Jeju Island, Korea, 2004.

[5] B. Meyer, T. Wesker, T. Brand, A. Mertins, and B. Kollmeier, A Human-Machine Comparison in Speech Recognition Based on a Logatome Corpus, *Proc. of ITRW on Speech Recognition and Intrinsic Variation*, Toulouse, France, 2006.

[6] E. Eide, H. Gish, P. Jeanrenaud and A. Mielke, Understanding and improving speech recognition performance through the use of diagnostic tools, *Proc. of 20th IEEE Conf. on Acoustics, Speech, and Signal Processing*, Detroit, USA, 1995, 221-224.

[7] L.L. Chase, Error-Responsive Feedback Mechanisms for Speech Recognizers, Ph.D. Thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, USA, April 1997. Also available as Robotics Institute Tech. Report # CMU-RI-TR-97-18.

[8] G. Greenberg and S. Chang, Linguistic dissection of switchboard-corpus automatic speech recognition systems, *Proc. of ITRW on Automatic Speech Recognition: Challenges for the new Millenium*, Paris, France, 2000, 195-202.

[9] T. Wesker, B. Meyer, K. Wagener, J. Anemüller, A. Mertins, and B. Kollmeier, Oldenburg Logatome Speech Corpus (OLLO) for Speech Recognition Experiments with Humans and Machines, *Proc. of Interspeech 2005*, Lisboa, Portugal, 2005, 1273-1276.

[10] P. Alku and T. Bäckström, Normalized aplitude quotioent for parametrization of the glottal flow, *J. of the Acoustical Society of the America*, 112(2), 2002, 701-711.

[11] M. Cernak and C. Wellekens, Emotional Aspects of Intrinsic Speech Variabilities in Automatic Speech Recognition, *Proc. of 11th Int. Conf. Speech and Computer*, Saint-Petersburg, Russia, 2006, 405-408.

[12] J.B. Allen, How do humans process and recognize speech?, *IEEE Trans. on Speech and Audio Processing*, 2(4), 1994, 567-577.

[13] M.M. Withgott and F.R.Chen, *Computational models of American speech*, Standford: Center for Study of Language and Information, 1993.

[14] R. Collobert, S. Bengio and J. Marithoz, Torch: a modular machine learning software library, IDIAP Tech. Report # IDIAP-RR 02-46, Martigny, Switzerland, 2002.

[15] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen, Classification And Regression Trees, New York: Chapman & Hall/CRC, 1983.

[16] J.B. Allen, *Articulation and Intelligibility*, New York: Morgan & Claypool, 2005.