ELSEVIER

Speech Communication 29 (1999) 81–82

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Editorial

The topic of this special issue is 'Modeling pronunciation variation for automatic speech recognition (ASR)'. The research interest in this topic is by no means new. Already more than 25 years ago, in April 1974, this topic was discussed at the 'IEEE Symposium on Speech Recognition'. Many articles in the proceedings of this symposium mention the need for including multiple pronunciations in the speech recognizer's lexicon and suggest the use of phonological rules to generate the various pronunciation forms. Recently, an increased interest in this topic has been observed. This renewed interest is probably a direct consequence of the shift in the type of speech used in ASR research: the type of speech used has gradually moved from isolated words to carefully read speech, and finally to speech that can be classified as extemporaneous or conversational, and is usually called spontaneous. It is clear that in going from isolated words to spontaneous speech the amount of pronunciation variation increases. Pronunciation variation is known to reduce the performance of ASR systems, if it is not well accounted for. Consequently, the larger amount of variation observed in spontaneous speech has made the need for pronunciation variation modeling even more urgent.

Since the IEEE symposium of April 1974, a lot of progress has been made in the field of ASR. The word error rates (WERs) for isolated words and carefully read speech have dropped below 10%. However, the WER for spontaneous speech is still much higher. In fact, it is too high for many practical applications. Modeling pronunciation variation is seen as a possible way of improving the performance of ASR systems that handle spontaneous speech.

Against this background, at the beginning of 1997 we decided to organize an international workshop on pronunciation variation modeling for ASR, so that researchers from different countries and language backgrounds – who work on this topic – could gather and discuss the problems they have encountered and try to find solutions to them. From 4 to 6 May 1998, the ESCA Tutorial and Research Workshop (ETRW) 'Modeling pronunciation variation for ASR' was held in Rolduc, Kerkrade, The Netherlands. In announcing the workshop to the research community, the topic was not further defined. This might seem strange at first, as one could argue that almost all ASR research is about modeling pronunciation variation. For instance, hidden Markov models (HMMs) are a way of modeling segmental and temporal variation, context-dependent HMMs are used to model the variation due to coarticulation effects, and multiple Gaussian mixtures can be employed to better model the segmental variation. Nevertheless, within the ASR community there seems to be an (undefined) notion of what falls under the heading of pronunciation variation modeling, as appears from a review of the papers presented under this heading at previous conferences, and from the papers submitted to this workshop. Apparently, the techniques mentioned above are no longer considered to be 'special techniques' for modeling pronunciation variation, but are more or less part and parcel of standard ASR. In general, when speaking about 'modeling pronunciation variation for ASR', one thinks about techniques other than the standard ones.

Of the submitted papers, 25 were selected for the workshop by the international scientific committee. In addition there were two invited speakers: Steve Greenberg and Michael Riley. In total, 72 participants from 20 countries in Europe, America, Asia and Africa took part in the workshop. As observed by Rolf Carlson in his report about the workshop which appeared in NESCA, the newsletter of ESCA, "The participants

PII: S0167-6393(99)00049-7

used this opportunity to ask questions and to make critical remarks and suggestions, making the event a true workshop with many active and engaged participants''.

At the beginning of the workshop evaluation forms were handed out to some of the participants, so that they could indicate the quality of the various contributions. Most of these evaluation forms were returned to us, either at the end of the workshop or later by post. On the basis of the returned evaluation forms, a selection was made from the 27 contributions to the workshop. Each of the selected papers was subsequently submitted to three reviewers. Eventually this resulted in nine papers being selected for this special issue. Not only were the selected papers qualified as the best ones in the evaluation forms, they also give a good overview of the different approaches used in this field.

What are the general conclusions that can be drawn from the workshop and the papers in this special issue? Relative reductions in the WER ranging from 0% to 20% have been reported so far. This can be interpreted positively: modeling pronunciation variation often reduces the WER, sometimes even by 20%. However, the general feeling at the Rolduc workshop seemed to be that the results so far did not live up to the expectations. Still, the research on modeling pronunciation variation has shown the importance of systematic lexicon design, and has yielded improved, more consistent lexica. Furthermore, different methods have been proposed and tested, and it has been shown that some are more successful in modeling pronunciation variation than others. Finally, it has become clear that modeling pronunciation variation is an important issue, and that the right solutions have not been found yet. Therefore, more research should be carried out. The key question is: What kind of research? Up till now, in the majority of the publications on this topic, different methods for modeling pronunciation variation were tested by simply comparing the WERs before and after the methods were applied. It has become clear that this is not enough. Not only should we try to find the methods that give the largest reductions in WER, we also should try to understand why this is the case and what kind of recognition errors are solved by these methods (i.e. what kind of pronunciation variation is modeled). Error analysis can be used for this purpose. Furthermore, more fundamental research is needed to unravel what kind of pronunciation variation is present in spontaneous speech, both in qualitative and in quantitative terms. After all, if we do not know what pronunciation variation really is, then how can we model it adequately?

Acknowledgements

I would like to thank the following persons, who all – in one way or another – contributed to this special issue. The international scientific committee of the workshop, mainly for reviewing all papers submitted for the workshop. The other members of the organizing committee of the workshop: Judith Kessens, Mirjam Wester, Febe de Wet, Loe Boves and Jean-Pierre Martens. The many people who returned the filled-in evaluation forms, which helped us in making a selection of the papers. Herve Bourlard, my 'contact' at *Speech Communication*, who always answered my questions swiftly and adequately. And last, but certainly not least, the reviewers of the papers of this special issue: Martine Adda, Louis ten Bosch, Loe Boves, Li Deng, Kjell Elenius, Sadaoki Furui, Finn Tore Johansen, Pat Keating, Sanjeev Khudanpur, Lori Lamel, Jean-Pierre Martens, Giorgo Micca, Roger Moore, Mari Ostendorf, Louis Pols, Simon Ringland, Elizabeth Shriberg, Torbjorn Svendsen and Klara Vicsi.

H. Strik
Department of Language and Speech, University of Nijmegen
A^2RT, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
E-mail: strik@let.kun.nl