Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 21 (1997), 155-157

SUMMARY OF A PH.D. THESIS DEFENDED IN 1997

Incorporating knowledge on segmental duration in HMM-based continuous speech recognition

author: Xue Wang
promotor: L.C.W. Pols
co-promotor: L.F.M. ten Bosch
date of defense: 14 April 1997

Automatic speech recognition (ASR) is a method for recognising spoken messages by computers. In the present-day state-of-the-art ASR, there are two competing approaches. The statistical approach based on HMM (hidden Markov model) currently outperforms the rule-based knowledge-engineering approach. Various attempts to combine the two approaches also exist. The research presented in this thesis took a viewpoint that in both of these approaches knowledge about speech is used, but this knowledge is represented in different ways. Our way to combine the two approaches is an attempt to incorporate specific knowledge into the HMM-based statistical ASR system. The technically feasible methods of knowledge incorporation were sought out in this thesis work, both based on the structure of HMM-based recogniser, and based on the complicated duration regularities in speech data.

In Chapter 1, first of all the current state-of-the-art in ASR is reviewed, leading to the conclusion that technical improvements are still necessary and possible for ASR. In our view the history of ASR development can be considered as a gradual process of incorporating specific knowledge about speech into the recognisers. Therefore each improvement generally implies the incorporation of a specific piece of knowledge. The current study concentrates on the knowledge about durational behaviour of the phonetic segments (phones), for reasons that there is a rich body of literature about this knowledge, and that the currently most successful HMM techniques have not appropriately incorporated this knowledge. Having chosen the HMM as the basic recogniser structure for this study, the problem of incorporating durational knowledge has two sides, namely on the one hand the durational behaviour of the HMM, and on the other hand the durational behaviour of the phonetic segments themselves as observed in the actual speech database. We were first of all confronted with the *linkage* problem between these two aspects, namely that there is no appropriate *representation form* of this knowledge, that can be used both to collect the knowledge from the database, and to incorporate the knowledge into the HMM-recogniser. This defines the general paradigm of the current study as a methodological one: searching for appropriate representations and searching for feasible ways of incorporation. Other technical specifications for the whole thesis work are also represented in this chapter, such as the use of monophone HMMs (for a manageable complexity and a tangible effect of duration modelling), and the (main) use of the TIMIT multi-speaker

continuous-speech database (this database being well documented and close to the situation of continuous speech recognition).

It was decided to incorporate knowledge about context independent (CI) and context dependent (CD) durations separately. The CI durational behaviour was investigated first. In order to find the relationship between the CI durational behaviour of the HMMs and the CI distribution of the segments, the mathematical basics of HMM are briefly reviewed in Chapter 2, together with a simple durational distribution for the case of a single state of HMM. General technical specifications for ASR research are reviewed, and the basic setups of the recogniser used in the current study are presented.

Based on the type of research of the current study, that was also defined in Chapter 1 as being technical, all different effects of knowledge incorporation should be tested in terms of the performance of a recogniser. Therefore a baseline system had to be built and optimised before any extra durational knowledge is incorporated. In Chapter 3 the optimisation was achieved by linear transformations of the front-end vectors to remove the correlation in them. Both filterbank parameters and mel-scale cepstral coefficients (MFCC) were tested, but ultimately MFCC plus their time-derivatives were chosen for the baseline system. Both a discrete- and a continuous-density system were tested, but only the latter was used in the rest of the thesis. The transformation on MFCC was performed using either the vectors in the whole database (global) or only the vectors assigned to an HMM state (state-specific). Slightly different impacts for phone and word recognition on the baseline performance with different transformation schemes were observed. These results are used for different purposes in the later chapters. The various implementations of linear transformations clarified the limitation of this technique in terms of its capability to improve the performance. This limitation exists mainly because the linear transformations that we used only removed the correlation, and do not improve significantly the modelling accuracy for non-Gaussian speech data in general.

Chapter 4 serves as a theoretical preparation for Chapter 5. In Chapter 4, the durational behaviour of the general left-to-right HMM is analysed. It is shown, with the help of theoretical durational pdf (probability density function), that even a linear HMM, as the simplest special case of a left-to-right model, is rich enough for modelling the single-peak binomial-like durational distributions of most phones. Therefore it is unnecessary to introduce for instance hidden semi-Markov model (HSMM) to repair the durational behaviour at the state level. Relations between the parameters of linear HMMs and the two lower statistics (duration mean and variance) of the phone segments were obtained, in order to be used in Chapter 5. The same relations for HMMs with skip transitions were also derived, but their complexity prevented their use in later chapters.

In Chapter 5, attempts were made to incorporate the CI durational knowledge (in the form of the CI duration mean and variance) into the HMMs. Several paradigms of training procedures were reviewed, including the one for HSMM, and the one for the standard HMM used by us with extra constraints on segmental durational statistics. The improved training procedure was embedded in the standard Baum-Welch maximum-likelihood (ML) framework. The durationally constrained ML equations were only solved numerically, giving duration fit for most of the phone HMMs in the system. This set of HMMs lead to better segmentation scores, indicating a better duration modelling accuracy. However, no (systematic) improvement in phone or word recognition was achieved.

CI durational modelling was thus considered to be insufficient, both because it did not lead to much improvement in system performance, and because the actual durational distribution is not context independent for sure. The influence of various

contextual factors on phone duration was systematically analysed in Chapter 6, to obtain context dependent (CD) durational knowledge. Durational distributions influenced by individual factors, as well as a nested ANOVA including all the 11 chosen factors, were used to reveal the CD durational behaviour. A number of factors appeared to be significant in influencing vowel duration, such as word stress, syllable locations within words and within utterances, and speaking rate, which would be used in the duration models for recognition in Chapter 7. The factor of voicing of post-vocalic stops did not show a systematic effect on the duration of the preceding vowels, thus this was not used in the next chapter.

In Chapter 7, four of the 11 contextual factors were chosen to be incorporated in the recogniser by means of a duration model that is external to the HMMs. Still monophone HMMs were used to generate the first *N* "*N*-best" sentence transcriptions for each utterance, at the word level. These word transcriptions were further used to generate phone-level transcriptions using the norm lexical pronunciation plus a word-juncture model. This model was derived from the same database and describes the pronunciation deviations from the norm at word junctures. The phone instances, with their duration and contexts identified, resulted in a phone duration score based on the duration model. The phone duration scores were integrated into the utterance duration score, and this was combined with the already available acoustic score of the *N*-best sentence transcription. The transcription with the highest combined score was taken as the new top-best. The word correct score of the new top-best transcriptions was marginally higher than the original top-best without this "re-scoring" process. In other words, the CD durational knowledge was incorporated into the recogniser in the post-processing phase.

The whole development of the current study indicates a possibility to incorporate statistically formalised knowledge on duration into a statistical recognition system based on a given structure of HMM. However the structure of this knowledge is defined by a phonetic parameter, being segmental duration. CI and CD types of durational knowledge were incorporated in different ways. The experience of the current study can be useful for incorporation of other long-term speech parameters (such as the pitch contour) into the frame-based HMM recognisers, which so far had been a difficult problem. Since the overall paradigm of the whole research project is new, a viewpoint on knowledge incorporation into machines was presented in Chapter 8, derived from our specific experience of incorporating durational knowledge, as a contribution to ASR research in general. This viewpoint takes an iterative process of knowledge representation in the form of "structure-plus-parameters".

The current study revealed that incorporating durational knowledge into HMM-based ASR is useful. To achieve this, however, in-depth analyses are required both on the structures presented in the speech data related to segmental durational behaviour, and on the structures of the given (HMM-based) recognition system in modelling various aspects including duration. Furthermore, the improvement of recognition performance will have to rely on careful engineering, both accomplished within this thesis work, and in possible future studies.