Text Independent Methods for Speech Segmentation

Anna Esposito¹,² and Guido Aversano²,³

¹ Seconda Università di Napoli, Dipartimento di Psicologia, Via Vivaldi 43, Caserta Italy  
anna.esposito@unina2.it, iiass.annaesp@tin.it  
² IIASS, Via Pellegrino 19, 84019, Vietri sul Mare, Italy, INFM Salerno, Italy  
³ École Nationale Supérieure des Télécommunications (ENST), 46 rue Barrault, 75634 Paris cedex 13, France  
aversano@tsi.enst.fr

Abstract. This paper describes several text independent speech segmentation methods. State-of-the-art applications and the prospected use of automatic speech segmentation techniques are presented, including the direct applicability of automatic segmentation in recognition, coding and speech corpora annotation, which is a central issue in today’s speech technology. Moreover, a novel parametric segmentation algorithm will be presented and performance will be evaluated by comparing its effectiveness against other text independent speech segmentation methods proposed in literature.

1 Introduction

The present work is situated in the context of non linear speech processing algorithms and in particular, it aims to present an overview of the proposed speech segmentation techniques. The main goal is to present a new algorithm, capable of performing phonetic segmentation of speech without prior knowledge of the phoneme sequence contained in the waveform.

Automatic segmentation of speech is such a difficult task that current-generation speech dictation and recognition systems avoid to deal with it directly; in such applications instead, statistical modeling of transitions between phonetic units seems to be preferable to an explicit, a priori fixed, segmentation.

Nevertheless, automatic language-independent segmentation is still an important challenge in speech research, playing a central role in today's speech technology, for applications like automatic speech recognition, coding, text to speech synthesis and corpora annotation.

Reliable automatic speech segmentation is of fundamental importance for Automatic Speech Recognition (ASR). In some way, the sequence of speech frames resulting from short-term analysis should be organized into homogeneous segments that are to be associated with a set of symbols representing phones, words, syllables, or other specific acoustic units. From this point of view, automatic speech segmentation could be highly useful, allowing to first segment the speech wave into a set of symbols and then implement the algorithms that recognize them. However, a technical solution to the segmentation problem that is completely satisfying (from the recognition point of view) has not been found yet [41]. This is one of the main reasons why the current state-of-the-art ASR systems do not operate in a bottom-up way, i.e. starting from an explicit segmentation. They instead adopt a generative, top-down approach, estimating the likelihood of top-level linguistic hypotheses, modeled as sto-
chastic processes, on the basis of the observed sequence of speech frames. Hidden Markov Models (HMMs) [74] are the most common implementation of the generative paradigm, and are widely used in both research and commercial applications of ASR systems. Nevertheless, the above approach tends to privilege lexical constraints over phonetic reality where known phonetic transcriptions are used to constrain the recognition of the speech signal into a forced alignment. A more balanced approach could be realized by including phonetic information into the generative decoding process. A possible step in this direction is to define a transition probability between speech units based on explicit segmentation and acoustic-phonetic cues [4].

The aim of speech coding applications is to generate a compressed representation of the speech signal for its efficient transmission or storage [43], [44]. Digital speech coding technologies are widely used in everyday-life applications, and in some of them, like cellular telephony, are of crucial importance.

Analysis methods are often shared between the ASR and coding areas. However, speech preprocessing and modeling for coding purposes must satisfy specific needs that are different from those of the ASR case. Actually, the speech signal has to be reconstructed from the coded information in all its variability and richness in order to preserve not only the linguistic message, but also many other physical features of the signal (that the listener can perceive).

Several speech coders are explicitly based on segmentation concepts such as the temporal decomposition segmentation method proposed by [3]. Moreover language-independent segmentation is essential to those techniques that realize very-low bit rate speech coding by indexing in a memory of speech segments [10], [24]. For these techniques, the segmentation of speech signals into phonetic units is crucial, since the incorrect detection of phone boundaries may significantly degrade the overall system performance. Therefore the development of more efficient speech segmentation techniques can largely help to improve speech coding techniques.

In speech research, the availability of large, annotated speech corpora is a fundamental issue. Such corpora are needed to train automatic speech recognition systems, and huge training data sets usually imply higher recognition accuracy [21]. Large quantities of labeled speech materials are also needed to obtain high-quality speech synthesis in text-to-speech (TTS) and coding applications. Collecting and manually annotating a speech corpus (either at word or phonetic level) is a complex and expensive task, especially for spontaneous speech recordings. As a consequence, reliable quantities of annotated speech data are available only for relatively few languages [41]. To overcome these difficulties several automatic techniques for phonetic transcription/annotation have been developed, regularly based on top-down ASR methods, like text-dependent Viterbi alignment with modeling of pronunciation variants (e.g. see [11], [12], [13], [16], [76], [92]). However, the above mentioned automatic annotation methods have good performances only if an accurate modeling is done of pronunciation variants and other phonetic phenomena like cross-word assimilation, elision, de-gemination, or dialectal variation, that are frequent in spontaneous speech. Hesitations, false-starts and other dysfluencies are another source of problem for these methods. In addition, the segmentation accuracy requirements for TTS applications are higher than those required for ASR applications, since ASR systems, being focused on the correct identification of the speech sequence, do not require, as the TTS systems, an accurate placement of phone boundaries. Text-independent segmen-