Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 347–364

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Application of continuous state Hidden Markov Models to a classical problem in speech recognition☆

Colin Champion*, S.M. Houghton

School of Electronic, Electrical and Systems Engineering, Gisbert Kapp Building, University of Birmingham, Edgbaston B15 2TT, United Kingdom

Received 2 April 2014; received in revised form 28 April 2015; accepted 5 May 2015
Available online 14 May 2015

Abstract

This paper describes an optimal algorithm using continuous state Hidden Markov Models for solving the *HMS decoding problem*, which is the problem of recovering an underlying sequence of phonetic units from measurements of smoothly varying acoustic features, thus inverting the speech generation process described by Holmes, Mattingly and Shearme in a well known paper (Speech synthesis by rule. Lang. Speech 7 (1964)).

Keywords: Speech recognition; Hidden Markov Model; Recognition by synthesis

1. Introduction

1.1. Overview

This paper addresses the problem of correctly incorporating dynamic information into the acoustic models used for speech recognition.

For several decades the dominant algorithms in the field have had recognised weaknesses in handling dynamics. The algorithms are based on Hidden Markov Models (*HMM*s) in which the state space is discrete – for this reason we will refer to them as Discrete State HMMs (or *DS-HMM*s). The feature vectors have been made up of spectral band energies or their transformation into cepstra. An overview of the use of HMMs in speech recognition is given by Gales and Young (2007).

The physical properties underlying speech consist of the smooth motion of articulators between positions defined by the various sounds. The same smoothness can be seen in acoustic features – at least for sonorant sounds – if they are chosen appropriately whereas features chosen for their ease of extraction may exhibit intractable dynamics.

Models of speech which preserve the salient features of the production process are attractive for use in recognition because they inherit the smoothness of the underlying mechanisms. Conventional recognisers fail in this respect for

☆ This paper has been recommended for acceptance by Shrikanth Narayanan.

* Corresponding author. Tel.: +44 121 414 2825.
E-mail address: c.champion@bham.ac.uk (C. Champion).

http://dx.doi.org/10.1016/j.csl.2015.05.001
0885-2308/

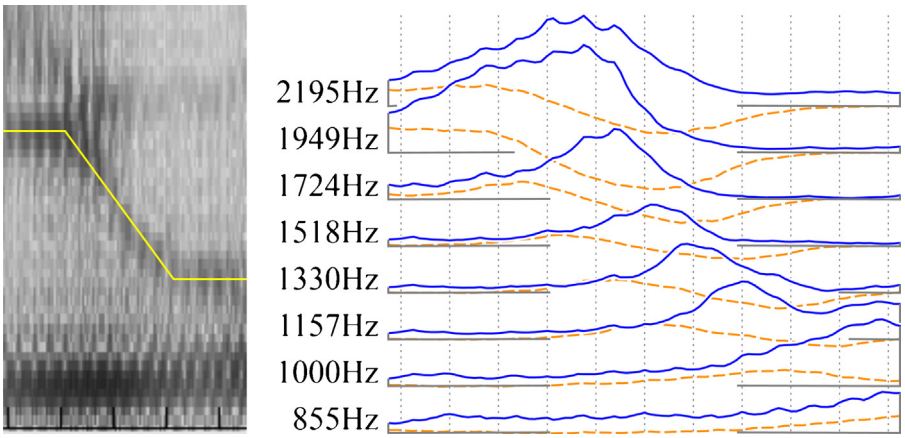

Fig. 1. Spectrogram showing the first two voiced phonemes of 'He will. . .' together with band energies and their derivatives for the spectral region containing F2.

two reasons. Firstly a continuous transition cannot be properly modelled as a sequence of discrete states, and secondly the approximately linear properties of the underlying features lose their structure when expressed in terms of spectral band energies.

The designers of speech recognition systems have sought to remedy these weaknesses by a number of strategies. One is to incorporate time derivatives (deltas) of cepstral features (Furui, 1981). This is easy to accomplish but questionable in terms of model coherence (Tokuda et al., 2003).

Fig. 1 shows the spectrogram for part of the TIMIT utterance 'He will allow a rare lie'. It contains most of the steady state (for which we will later define the term *dwell*) of the 'e' of 'he', the transition to the following 'w', and most of the corresponding steady state. The F2 path (idealised as a piecewise linear yellow line) is fully intelligible in terms of the phonetic sequence. The solid blue lines on the right of the figure show the energies of triangular mel-spaced bands in the F2 region. Each of these makes sense in terms of the formant motion but not phonetically; it seems odd, for instance, to say that the transition from 'e' to 'w' is characterised by a brief activation of the band at 1724 Hz about an eighth of the way through. The linearity of the formant feature has been replaced a complicated interdependence between band energies.

The orange dashed lines are the derivatives of the band energies (computed in the normal way).

The second strategy for capturing dynamic properties in a speech recogniser is to add parameters whose function is to describe transitions explicitly. We contend that it is more natural and more economical to infer the properties of transitions through the structure of the model. This is the subject of the present paper, in which we seek to show how models of individual phonetic units can be constructed in such a way that the properties of transitions can be inferred by interpolation. A model with these properties is likely to be closely related to the speech production process (as ours is), since acoustic features gain their interpolability by reflecting the smooth motion of articulators.

Standard DS-HMMs try to model transitions by splitting each sound into substates some of which fall in transition regions, and then estimating the parameters of each phonetic unit according to its context. If the transition from φ_0 to φ_1 is split into sufficiently many substates, and if the properties of each substate are estimated for the pair (φ_0, φ_1) rather than for an individual phonetic unit, then a good enough approximation will be obtained.

The main penalty to capturing properties through parameters rather than through model structure is that the amount of training data needed by an algorithm increases with the size of its parameter space. But a more insidious penalty lies in the fact that an algorithm whose strength lies in the number of its parameters sheds no light on the problem it addresses.

1.2. Views of speech dynamics

A number of attempts have been made to incorporate faithful models of speech dynamics into recognition algorithms. Many of these have been based on segment models whose theory was developed by Gales and Young (1993), Russell (1993), Holmes and Russell (1999), and others.