Structural Representation and Matching of Articulatory Speech Structures based on the Evolving Transformation System (ETS) Formalism

Alexander Gutkin
Centre for Speech Technology Research
University of Edinburgh
Edinburgh, EH8 9LW, UK
alexander.gutkin@ed.ac.uk

David Gay
Faculty of Computer Science
University of New Brunswick
Fredericton, NB, E3B 5A3, Canada
dave.gay@unb.ca

Abstract

A formal structural representation of speech is presented in this paper. The representation is developed within the Evolving Transformation System (ETS) formalism and encapsulates speech processes at the articulatory level. We show how the *class structure* of several consonantal phonemes of English can be expressed via articulatory gestures. Elements of these classes can be detected in a corresponding *structural representation* of continuous speech. Our experimental results on the MOCHA articulatory corpus support the hypothesis that the proposed articulatory representation captures sufficient information for the accurate structural identification of phonemic classes.

1 Introduction

Despite the evident success of numerical approaches to modelling of speech, they are often criticised for having little relation to actual human speech production/recognition [Jelinek, 1997, p. 10]. The alternative, structural, means of pattern representation have, however, received little attention. In our view, one of the main reasons for this situation is the apparent lack of suitable structural frameworks possessing the necessary formal power to accommodate the representation of classes of complex linguistic phenomena (e.g. phonemes and syllables). It is hypothesised that the appearance of such a systematic analytical framework and the development of appropriate representations within it may help to bridge the gap between speech recognition and linguistic research.

On the linguistic side, the motivation for this work comes from the theory of articulatory phonology [Browman and Goldstein, 1992], in which vocal tract action during speech production is decomposed into discrete, recombinable atomic units known as *gestures*. Compared to traditional approaches—such as distinctive phonological features [Jakobson and Halle, 1971]—the gestural approach is more physiologically concrete and offers a compact means of representing the truly asynchronous nature of speech, allowing for better interpretations of complex phonological phenomena (such as co-articulation).

The Evolving Transformation System (ETS), outlined in [Goldfarb *et al.*, 2004], is a radically new formal framework for the structural representation of "natural" processes. ETS suggests that the representation of each such process should include its "formative history", which is a series of "operations" (*primitives*) acting on the constituent elements (*sites*) of the process. Such formative histories (*structs*) should contain some regular "chunks" (*transforms*). A class representation in ETS is a finite set of closely related transforms, out of which the corresponding class elements—processes—can be constructed. In our ETS representation of gestural speech structure, the natural process we model is the physiological process of articulation.

Organ	Semantics
UL	upper lip
LL	lower lip
UI	upper incisor
TD	tongue dorsum
TT	tongue tip
VL	velum
HP	hard palate
AR	alveolar ridge
VF	vocal folds

Group	Organs	G
bilabial closure	UL, LL	6
tongue dorsum height	TD	4
tongue tip height	TT	4
labiodental contact	UI, LL	4
velic aperture	VL	4
velar contact	TD, VL	2
alveolar contact	TT, AR	2
palatal contact	TT, HP	2
voicing	VF	2

Table 1: Articulatory organs (left) involved in the production of various groups of primitive gestures (right).

2 Articulatory Structure according to ETS

The units of representation corresponding to articulatory organs and primitive articulatory gestures are ETS *sites* and *primitives*, respectively. Informally, an ETS primitive is a unit of temporal structure of a process, describing the structural transformation of its set of "initial" sites into its set of "terminal" sites, where an ETS site is the smallest/unstructured representational unit within a process. In the articulatory representation, a primitive is identified with a change in the interaction of one or more of the associated articulatory organs, which are expressed as sites.

The left-hand side of Table 1 lists all of the ETS sites used in the representation, along with the corresponding interpretation. The right-hand side of Table 1 shows the groups of primitive gestures used in this study. For each group, the relevant sites (articulators) and the number of distinct constituent gestures (primitives) G are shown. Informally, a group consists of semantically and syntactically related primitive gestures involving similar articulators. Pictorially, it is convenient to

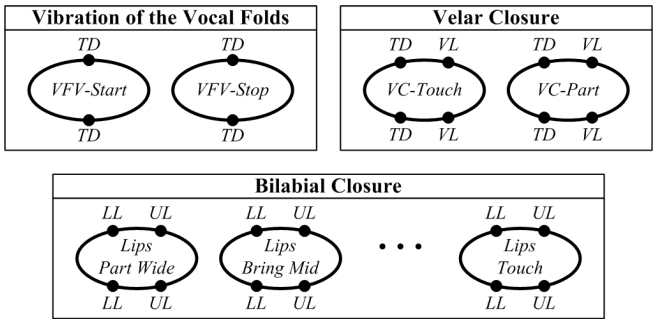

Figure 1: Several of the ETS primitives used in this representation, grouped by articulatory category.

represent primitives as convex shapes, with initial sites depicted as points on the upper half, and terminal sites depicted as points on the lower half. Some of the 30 primitives used in the articulatory representation are given in Figure 1. Three groups are shown: two articulatory gestures of the vocal folds resulting in voiced or unvoiced sounds, the two gestures participating in velar closure, and three of the six gestures modelling the aperture of the lips (bilabial closure).

An ETS *struct* is a temporally-ordered sequence of connected primitives capturing the "history" of the corresponding process. Within our articulatory representation, a struct is identified with a sequence of primitive gestures, which are hypothesised to encapsulate the gestural structure of any given utterance (note that any utterance can itself be interpreted as a highly non-trivial gesture). Figure 2 depicts the gestural structure of the word "get" in our representation. The articulation of $/g/$, for instance, has a simple interpretation: a velar constriction is formed and then released. The formation of the constriction is achieved when the tongue dorsum TD first rises to its maximum position (TD-RaiseMax) at 0.248 sec completing the constriction before the phoneme boundary by touching the velum VL (VC-Touch) at 0.266 sec). The constriction release is accomplished by lowering the tongue dorsum TD (TD-LowerMid) at 0.416 sec, then parting the tongue dorsum TD from the velum VL (VC-Part) at 0.460 sec). Note: vibration of the vocal folds VF (VFV-Start) occurs at the onset of $/g/$ at 0.380 sec.

An ETS *transform* is an encapsulation of a regular temporal pattern of primitives. An ETS *supertransform* is a set of closely-related transforms specifying the description of a class, where structural variations account for noise in the class. Note that the learning of class structure is described in [Goldfarb *et al.*, 2004] and is outside of the scope of this paper. In the articulatory representation, a supertransform is identified with the family of articulatory gestures that collectively describe the class structure of a single, general phoneme. We defined 14 consonantal classes (see, for example, the identified phoneme $/g/$ in Figure 2).

We developed a *structural matching algorithm* for detecting the presence of an ETS transform in a given struct. It performs a rooted depth-first search in $O(n^2 \log n)$ time, where n is the number of primitives in the transform to be matched.

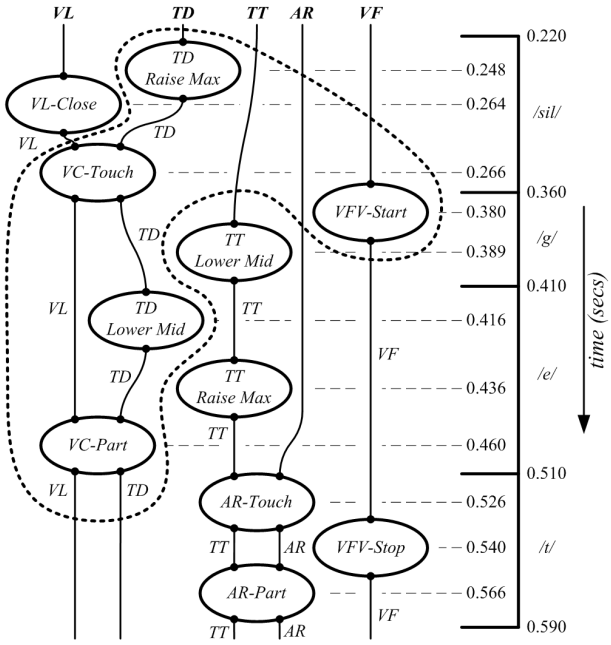

Figure 2: Partial ETS struct describing the gestural structure of the word "get", constructed using primitive gestures detected on-the-fly in sample data. The transform corresponding to phoneme $/g/$ is identified with a dotted line.

3 Experiments and Discussion

Our goal was to assess the performance of the 14 ETS supertransforms via the structural matching algorithm against structures derived following [Gutkin and King, 2005] from real data (920 MOCHA utterances, 32169 ETS primitives).

Out of 9878 phonemes, 7679 were classified correctly and 278 failed to match against any of the 14 available classes. The overall accuracy is 77.74%. Analysis shows that $/n/$ was often misclassified as $/d/$ and $/m/$ as $/b/$, both due to a failure of the pre-processor to detect the corresponding physiological changes. We expect performance to improve with a more accurate pre-processor and refined class descriptions.

References

[Browman and Goldstein, 1992] C. Browman and L. Goldstein. Articulatory Phonology: An Overview. *Phonetica*, 49:155–180, 1992.

[Goldfarb *et al.*, 2004] L. Goldfarb, D. Gay, O. Golubitsky, and D. Korkin. What is a structural representation? Technical Report TR04-165, Faculty of Computer Science, University of New Brunswick, Canada, April 2004.

[Gutkin and King, 2005] A. Gutkin and S. King. Detection of Symbolic Gestural Events in Articulatory Data for Use in Structural Representations of Continuous Speech. In *Proc. ICASSP*, Philadelphia, March 2005.

[Jakobson and Halle, 1971] R. Jakobson and M. Halle. *Fundamentals of Language*. Mouton de Gruyter, 1971.

[Jelinek, 1997] F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA, March 1997.