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

Alexander Gutkin
School of Informatics
University of Edinburgh
Edinburgh, EH8 9LW, UK
alexander.gutkin@ed.ac.uk

David R. 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 consistent with the principles of combinatorial structure theory 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 with the help of articulatory gestures — the atomic combinatorial qualitative units of speech. As a preliminary step towards the design of a speech recognition architecture based on the structural approaches to physiology and articulatory phonology, we present an algorithm for the structural detection of phonemic class elements inside gestural ETS structures derived from continuous speech. Experiments designed to verify the adequacy of the hypothesised gestural class structure conducted on the MOCHA articulatory corpus are then described. Our experimental results support the hypothesis that the articulatory representation captures sufficient information for the accurate structural identification of the phonemic classes in question.

1 Introduction

We begin with the observation that, although the traditional means of studying speech phenomena in linguistics have been structural, the approaches to pattern representation in speech recognition are predominantly numerical. Despite the evident success of numerical approaches, especially in the area of automatic transcription, 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 qualitative class representation of complex linguistic phenomena (e.g. phonemes and syllables). Sadly enough, this state of affairs also appears to apply to many other areas of pattern recognition [Pavlidis, 2003, Section 3]. It is hypothesised that the appearance of such a systematic analytical framework and the development of appropriate qualitative representations within it could potentially help in bridging the gap between, in particular, 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 articulatory phonology, vocal tract action during speech production is decomposed into discrete, re-combinable atomic units. The central idea is that while the observed products of articulation (articulatory and acoustic measurements) are continuous and context-dependent, the physiological actions which engage the organs of the vocal tract and regulate the motion of the articulators are discrete and context-independent. These atomic actions, known as *gestures*, are hypothesised to combine in different ways to form the vast array of words that constitute the vocabularies of human languages [Kennedy and Goldstein, 2003]. This combinatorial outlook on speech places it in the same context as other natural systems (for instance, combinations of gestures are similar to molecular compounds in chemistry). 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 all-pervasive complex phonological phenomena (such as co-articulation). In Section 2 we briefly discuss several issues in qualitative gestural representation of the speech production process.

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 process's constituent elements (*sites*). Such formative histories (*structs*) should contain some regular "chunks" (*transforms*), which are the building blocks of class representation.

A class representation in ETS is a finite set of closely related transforms (a *supertransform*), 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. This speech representation is introduced in Sec-

Alexander Gutkin and David R. Gay, *Structural Representation and Matching of Articulatory Speech Structures based on the Evolving Transformation System (ETS) Formalism*, Proc. 19th International Workshop on Qualitative Reasoning (QR-05) (Graz, Austria) (Michael Hofbaur, Bernhard Rinner, and Franz Wotawa, eds.), May 2005, pp. 89–96.

tion 3 where, on the articulatory level, the gestural structure of speech is captured by fundamental concepts of the ETS formalism, such as sites and primitives. The gestural structure of various classes of consonantal phonemes of English is also described.

The purely quantitative approach to automatic derivation of gestural structures from articulatory speech data has been studied in detail in [Jung *et al.*, 1996], where the authors proposed using a derived numeric representation of the gestural structure both as alternative units for continuous speech recognition and as a compact representation of the acoustic waveforms. An alternative (qualitative) approach, advocating the use of automatically derived gestures as the generic qualitative units for any *structural* representation of continuous speech, is presented in [Gutkin and King, 2005]. In this paper, the latter approach (briefly outlined in Section 5) was used for detection of the articulatory gestures in the continuous speech data and automatic derivation of gestural structures for ETS representation. Within ETS, symbolic articulatory events have a natural interpretation: the articulators are identified with ETS sites, while the elementary gestures are identified with ETS primitives.

Given an ETS-based class representation of phonemes[1], provided in terms of gestural structure, a structural matching algorithm for the detection of class elements in articulatory structures derived from real data is presented in Section 4. Experiments aimed at verifying the accuracy of identification of the 14 consonantal phonemes of English using the above structural matching are described in Section 5. These experiments were conducted on the entire MOCHA corpus, described in [Wrench, 2000], which consists of articulatory recordings of continuous speech. We conclude the paper with Section 6, where we present the potential benefits of this approach to the speech science community.

2 Qualitative Articulatory Modelling

As mentioned by Forbus [1996], "there is no single, universal right or best qualitative representation. Instead there is a spectrum of choices, each with their own advantages and disadvantages for particular tasks". In this section, we briefly outline some of the modelling assumptions made when describing the continuous non-linear properties of the speech production process.

2.1 Level of Resolution

The theoretical motivation for the choice of the articulatory level of resolution is supported by linguistic theory, which states that an analysis on a lower, motor, level [Zemlin, 1968; Kaplan, 1971] introduces too much anatomical detail which is linguistically irrelevant for the discrimination between various sound patterns [Ladefoged, 2001]. Articulatory gestures can therefore be seen as optimal representational units, which on one hand have direct physiological reality, and on the other provide the qualitative combinatorial structure for various linguistic classes. In the ETS formalism, which we use for modelling, an articulatory gesture is described by an ETS primitive, the smallest structured unit of the model encapsulating both syntax and semantics.

[1]The process of learning a class structure, outlined in [Goldfarb *et al.*, 2004], is outside the scope of this work.

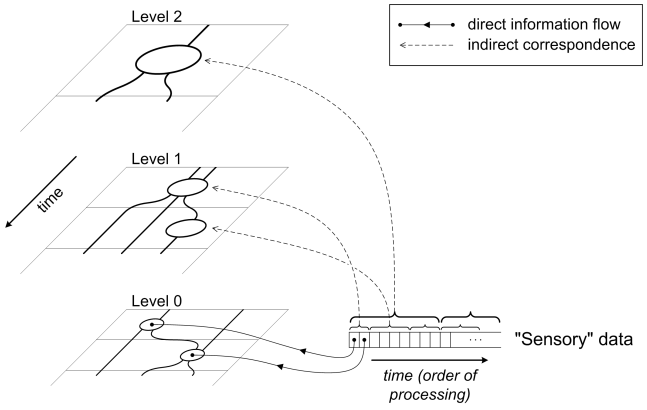

Figure 1: A multi-level ETS representational tower with a single-level sensor at level 0. Reproduced from [Goldfarb *et al.*, 2004].

2.2 Compositionality

One of the definitions of the compositionality, adopted in neuroscience [Abeles *et al.*, 2004], states that it is the ability to construct mental representations *hierarchically*, in terms of parts and their relations. The notion of hierarchical mental representations is of paramount importance in the study of speech and language.

The ETS formalism has been designed to support the view of the environment as formed by a hierarchically evolved — and therefore hierarchically organised — dynamic system of classes (of entities). An ETS view of environment is therefore *multi-leveled*, with the processes at each level composed of the ETS primitives, each of which (except for the initial level), stands for a class from the previous level [Goldfarb *et al.*, 2004; Goldfarb, 2004]. In ETS, the level of abstraction increases as one ascends the levels in the hierarchy. The formalism provides the same formal language for dealing with each level of the representation. A useful metaphor for capturing the hierarchical representation is a multi-level "evolving representational tower" which can "sense" (and interact with) data *only* at the initial level (see Figure 1). In this paper we only deal with the bottommost, articulatory, level of representation which corresponds to the sensory level in the hierarchy.

At the articulatory level of representation, the ETS primitives (gesture) combine together in different ways to form the formative histories (gestural structures) of the speech production process. Semantics of the gestural structure is therefore determined on the basis of its constituent primitives (some of the examples are given in Section 3).

3 Articulatory Structure according to ETS

3.1 Gestures as ETS Primitives

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

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). See also Figure 2.

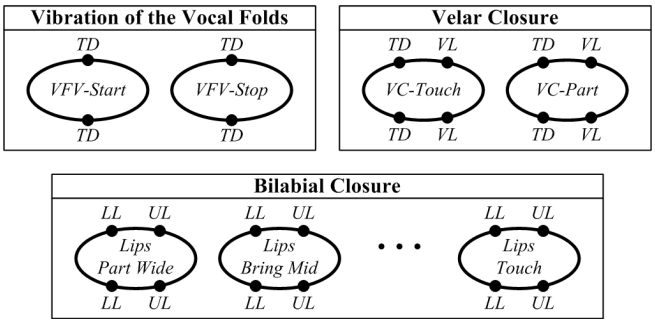

Figure 2: Several of the ETS primitives used in this representation, grouped by articulatory category.

of temporal structure of a process, describing the structural unit that transforms 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 ETS, a primitive is defined in such a way that its syntax and semantics are inseparable [Goldfarb, 2004].

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. An important simplifying assumption made in this study is that the sets specifying the initial and terminal sites of each primitive are identical since the number and type of the articulatory organs involved in the production of any given gesture do not change with time.

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.

Example. Pictorially, it is convenient to 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 presented in Figure 2. 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).

3.2 Gestural Formations as ETS Structs

An ETS *struct* is a temporally-ordered sequence of connected primitives capturing the "history" of the corresponding process. Within an articulatory representation, a struct is identified with a temporal sequence of primitive gestures, which in this paper are hypothesised to provide the gestural structure of any given utterance. We note that any utterance can itself be interpreted as a highly non-trivial gesture.

Example. Figure 3 shows an ETS-based gestural structure of the word "get", consisting of 11 primitive gestures operating on 5 articulators, together with the corresponding phonetic segments (detection and construction processes do not make use of these segments). Names of all the articulators (corresponding to ETS site types) used in this work are given in Table 1. The gestural structure in Figure 3 is constructed on-the-fly from the primitive gestures detected in the available articulatory and acoustic data. For the sake of clarity, only some of the primitive gestures participating in the critical articulation of the voiced velar stop $/g/$ and the unvoiced alveolar stop $/t/$ are shown. The articulation of $/g/$, for instance, has a simple interpretation within this representation. Articulation is achieved by first forming a velar constriction, which, in turn, is formed by the tongue dorsum TD first rising to its maximum position (TD-RaiseMax) at 0.248 sec, then completing the constriction before the phoneme boundary by touching the velum VL (VC-Touch) at 0.266 sec. The constriction is released within the phoneme boundaries of $/e/$ by first slightly lowering the tongue dorsum TD (TD-LowerMid) at 0.416 sec and then parting the tongue dorsum TD from the velum VL (VC-Part) at 0.460 sec. Note that vibration of the vocal folds VF (VFV-Start) occurs at the onset of $/g/$ at 0.380 sec. Similarly, it is possible to analyse the unvoiced alveolar stop $/t/$, the articulation of which is obtained by means of the tongue tip (TT), alveolar ridge (AR), and the vocal folds (VF).

3.3 Class Description via ETS Transforms

An ETS *transform* is an encapsulation of a regular temporal pattern of primitives, which is subdivided into two parts: the *context* and the *body*. The context of a transform identifies the place, within a given struct, in which the application of the body of the transform becomes legal, while the body is the "chunk" that extends the (previously constructed) struct. For the purposes of the articulatory representation of speech, a transform's context ensures that all of the gestures required for the formation of the constriction at the onset of the phoneme have occurred, while the body of a gestural transform specifies the release of the constriction. Hence, the phoneme starts with the first primitive in the context of the corresponding transform and ends with the last primitive in the body.

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. In the articulatory representation, a supertransform is identified with the family

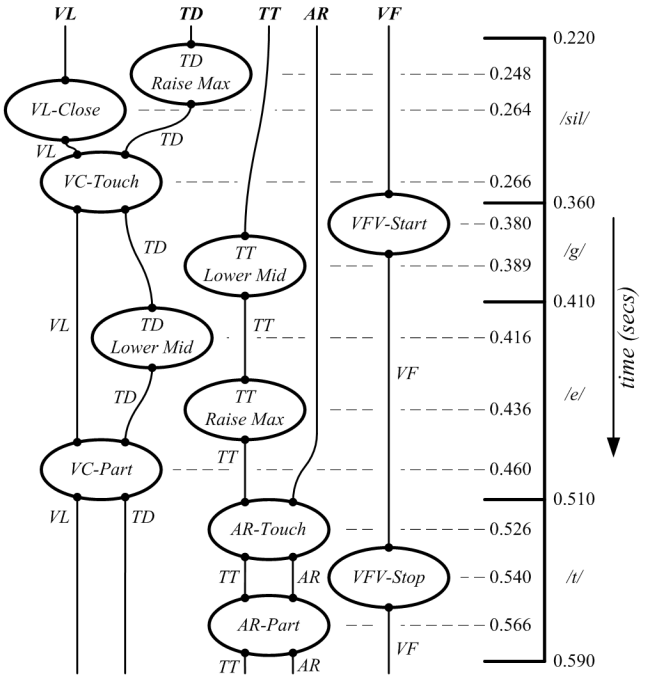

Figure 3: ETS struct describing the gestural structure of the word "get", constructed using primitive gestures detected automatically in sample data. Corresponding phonetic labels are shown.

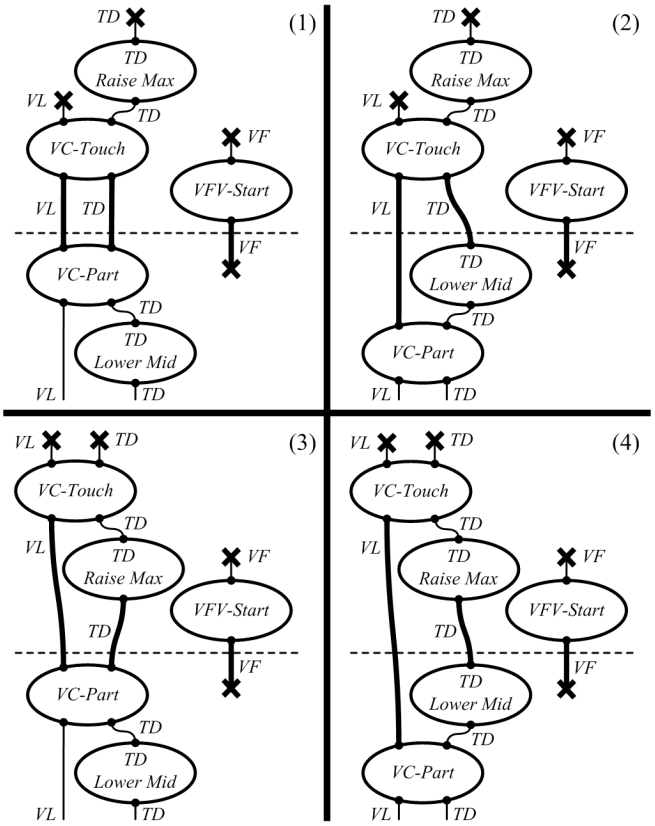

Figure 4: Simplified depiction of an ETS supertransform for the class of voiced velar stops given by phoneme $/g/$, consisting of four ETS transforms.

of temporal patterns of articulatory gestures that collectively describe the class structure of a single, general phoneme.

Example. A simple class structure for the class of voiced velar stops defining phoneme $/g/$ is given in Figure 4 (phoneme labels and timestamps corresponding to primitives are not shown). Note that the constituent transform (2) corresponds to the particular instance of the articulation of $/g/$ shown in Figure 3. Each column of the supertransform representation consists of transforms which have structurally identical bodies (with each body specifying a release of constriction). The thicker lines connecting constituent gestures between the body and context of each constituent transform denote interface sites — used to indicate that the necessary precondition (provided by the context) for the articulation of the respective phoneme has been met.

In this work, we have chosen to focus on class elements which are provided in terms of contexts of the gestural transforms. Our reason for doing this is that, for any given gestural transform, the detection of the context alone is enough to decide whether the phoneme corresponding to that transform has occurred. In addition, because phonetic labels have been provided with the data in this study, there is no need for duration modelling (for which one needs both body and context information). Henceforth, when referring to the gestural structure of phonemes, constituent gestural transforms are assumed to consist of contexts only.

Based on linguistic evidence [Ladefoged, 2001], only some of the primitive gestures from the gestural groups given in Table 1 were postulated to be critical for structural description of each of the 14 consonantal phonemes evaluated in this study. These phonemes, together with the names of the constituent primitive gestures, are shown in Table 2. For example, the gestural structure of the unvoiced alveolar stop $/t/$, is specified by a supertransform having six distinct constituent transforms (not shown), each consisting of various combinations of the three gestures VFV-Stop, AR-Touch, and TT-RaiseMax.

4 Matching Articulatory Transforms

It is desirable to have a procedure, called a *structural matching algorithm*, for detecting the presence of an ETS transform in a given struct. In particular, we only need to consider the case when the addition of a primitive to the end of a struct (i.e. as the preprocessing front-end generates gestural events) causes the "completion of construction" of a transform. Thus, the algorithmic approach presented here may be seen as a rooted depth-first search [Valiente, 2002], commencing with the last primitive in the body of a transform (corresponding to the latest primitive in the struct to be searched).

In order to reduce the number of structurally-equivalent transforms in a supertransform, a generalization of the specification of transform context/bodies to partial orderings of primitives — rather than total orderings — is possible (see Figure 5). This approach complicates structural matching,

Phoneme	Names of Constituent Gestures
/b/	VFV-Start LipsTouch VC-Part AR-Part HP-Part
/p/	VFV-Stop LipsTouch VC-Part HP-Part
/g/	VFV-Start VC-Touch TD-RaiseMax AR-Part HP-Part
/k/	VFV-Stop VC-Touch TD-RaiseMax AR-Part HP-Part
/d/	VFV-Start AR-Touch TT-RaiseMax
/t/	VFV-Stop AR-Touch TT-RaiseMax
/v/	VFV-Start LD-Touch
/f/	VFV-Stop LD-Touch
/ng/	VFV-Start VC-Touch TD-RaiseMax VL-Close*
/m/	VFV-Start LipsTouch VL-Close*
/n/	VFV-Start AR-Touch TT-RaiseMax VL-Close*
/ch/	VFV-Stop TT-RaiseMax AR-Touch
/zh/	VFV-Start TT-RaiseMax HP-Touch
/sh/	VFV-Stop TT-RaiseMax HP-Touch

Table 2: Phonemes and names of the constituent gestures participating in the formation of the respective articulatory constrictions. The name VL-Close* denotes *any* of the velic aperture gestures resulting in any degree of velum opening, excluding closure.

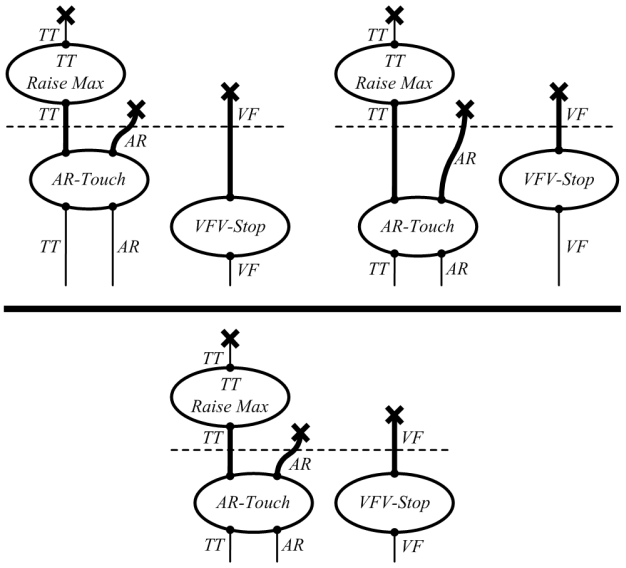

Figure 5: Top: Pictorial representation of two distinct transform fragments. Bottom: Representation of the above fragments as a single fragment, w.r.t. the partial order specification. Note that the "bottommost primitives" are AR-Touch and VFV-Stop (neither is attached to any succeeding primitive).

though a proof-of-concept implementation has been successfully developed by the second author. A pseudocode version of this algorithm is presented in Figure 6.

1. Let Π denote the (ordered) set of n primitives in the struct to be matched and Γ denote the (partially ordered) set of m primitives in the transform to be matched.
2. For each "bottommost" primitive $\gamma_i \in \Gamma$ (see Figure 5) that is of the same type as π_n, $\pi_n \in \Pi$, perform the following search:
   Declarations
   ρ: a *current primitive* ($\rho \in \Gamma$)
   V: a set of *visited* γ*-primitives*
   P: a set of *pending* γ*-primitives*
   E: an *equivalent primitive* mapping $E\colon \Gamma \rightarrow \Pi$
   $V \leftarrow \varnothing$
   $P \leftarrow \{\gamma_i\}$
   $E \leftarrow \{\gamma_i \mapsto \pi_n\}$
   `WHILE` $P \neq \varnothing$,
   $\rho \leftarrow P.\mathrm{Pop}()$
   `IF` $\rho \in V$, `CONTINUE`
   `FOR` each primitive α attached to ρ,
   Let β be the corresponding primitive attached to $E(\rho)$
   `IF` $\alpha \in V$, `NEXT`
   `IF` $\mathrm{Type}(\alpha) \neq \mathrm{Type}(\beta)$
   Try next γ_i (return to step 2)
   `IF` $E(\alpha)$ exists `AND` $E(\alpha) \neq \beta$
   Try next γ_i (return to step 2)
   `IF` $E(\alpha)$ does not exist
   $E \leftarrow E \cup \{\alpha \mapsto \beta\}$
   $P.\mathrm{Push}(\alpha)$
   $V.\mathrm{Push}(\rho)$
   `HALT:ACCEPT`
3. `HALT:REJECT`

Figure 6: ETS Transform Matching Algorithm.

A transform is accepted when the structure of the transform is detected inside the searched struct, i.e. when all primitives in Γ are mapped to primitives of the same type and "interconnectedness" in Π. A mismatch in type or interconnection causes the rejection of a transform. The worst-case complexity of this algorithm is $O(m^2 \log m)$, where m is the number of primitives in the transform to be matched.

All of the primitives used in this representation are divided into various articulatory groups, each consisting of semantically and structurally related primitives. Additionally, all m constituent primitives $\gamma \in \Gamma$ belong to separate groups (see Table 2). This allows for the trivial modification of the matching procedure specified above, whereby any primitive gesture $\pi \in \Pi$ which does not belong to any of the m groups is skipped during the search (without aborting the search procedure). This modification allows the detection of candidate transforms in those cases when a "structural overlap" of various class elements appears in the data. We used this modification of the search algorithm in our work.

5 Experiments

The aim of the experiments described below was to assess the performance of the structural identification of the 14 ETS

supertransforms — each describing a consonantal phoneme of English (see Table 2) — in gestural ETS structures derived from real articulatory data.

The articulatory corpus used for the experiments was the MOCHA corpus, consisting of articulatory and acoustic recordings of 460 phonetically-rich sentences designed to provide good phonetic coverage of English. The database contains finalised recordings of one male (acronym `msak`) and one female (acronym `fsew`) speaker of British English, each consisting of approximately 31 minutes of speech. Data channels include electromagnetic articulograph (EMA) sensors providing information about the trajectories of the articulators, a laryngograph/EGG channel measuring changes in the contact area of the vocal folds, and electropalatograph (EPG) measurements providing tongue-palate contact data. For more information on the data provided by MOCHA, refer to [Wrench, 2000].

Structural matching experiments were conducted on the combined dataset containing the recordings for both male and female speakers. The combined dataset consists of 920 sentences with an overall duration of one hour and two minutes. An algorithm for the automatic detection of symbolic articulatory gestures in the articulatory data, proposed in [Gutkin and King, 2005] (the overall error made by the pre-processor in detecting the symbolic gestures on the same dataset was 7.29% for the female speaker and 8.17% for the male speaker), was used to derive a gestural structure for each of the utterances in the corpus. The gestural structures make use of the 30 primitive gestures from the nine groups given on the right-hand side of Table 1 and are comprised of 32,169 primitives in total.

5.1 Detection of Primitive Gestures

In this section, we briefly summarise the mechanism of mapping from the various numeric measurements provided by MOCHA, to the basic qualitative units of the model – ETS primitives (shown in Table 2). The following exposition is based on [Gutkin and King, 2005].

Vibration of the vocal folds (VF) that uniquely defines voiced and unvoiced sound patterns is represented by the two primitives standing for the beginning (VFV-Start) and end (VFV-Stop) of vibration respectively. The pitch detection algorithm used on the acoustic recordings provided by the MOCHA database is described in [Talkin, 1995]. We used a 5 ms interval for analysis frames and a pitch frequency search range between 25 Hz and 600 Hz. Given the acoustic stream, at any given point in time the decision about the beginning and termination of the vibration is made when a change in the state of pitch is detected by the pitch detection algorithm, provided this new state is steady for at least 20 ms (around 320 samples of a 16 kHz recording), which is an average duration of a typical short vowel.

Given the EPG stream provided by MOCHA it is possible to detect various contacts between the tongue and the hard palate. The output of the EPG sensor consists of 8 8-bit binary vectors with a simple spatial structure. The first three rows represent the alveolar region (the first and the last bit of the first row are unused), followed by two rows representing the palatal region, with the last three rows roughly corresponding to the velar region. In order to determine whether contact has occurred, for each of the three regions (velar, palatal and alveolar) we use the contact index measured by the linear combination of the rows representing that region (which is a sum of all the bits of the rows), as described in [Nguyen, 2000]. Given an appropriate per-region threshold (τ_a,τ_p and τ_v representing the alveolar, palatal and velar regions, respectively) defined by examining the relevant EPG measurements, change in the contact information at any given point results in the emergence of an appropriate primitive if and only if the threshold value of the index is crossed. For instance, the velar contact gesture VC-Touch emerges when the value of the velar index increases beyond τ_v, while the gesture VC-Part signifying the release of the closure emerges when this value decreases below τ_v. The emerging primitive gestures involve the pair of organs corresponding to the contact location. For palatal contact, the organs would involve tongue tip (TT) and the hard palate (HP), for alveolar contact the pair would include the tongue tip (TT) and the alveolar ridge (AR). Since the EPG sampling frequency of 200 Hz is reasonably low and the measurements appear to change slowly over time, we have not imposed any requirements on the values of the indexes to be steady for any period of time.

The data stream containing EMA trajectories provides additional information about the articulations. Since the primitive gestures to be detected in the EMA data have a discrete nature, an obvious approach we follow is to cluster the distance measurements between the pair of the articulators of interest. The clustering, making use of an efficient variant of k-means described in [Kanungo *et al.*, 2002], is applied to the entire data available for the particular speaker. Since the vocal tract configurations vary from speaker to speaker, the clustering procedure is speaker-dependent. Each of the n cluster centroids represents one of the n regions of the vocal tract. For any given EMA frame, the distance between the two articulators is calculated and compared to the nearest cluster centroid. If the nearest centroid for this pair of articulators has changed since the last frame and the current articulation is sustained for at least m frames, the decision is made to fire an ETS primitive which represents the event responsible for a change in the state of the articulation. We consider the articulation to be sustained for m frames if the measurements of the distances between the two articulators for each of the m frames fall into the same cluster.

If a single articulator is involved in a gesture (for instance, the gesture TT-LowerMid only involves one articulator), the height of the articulator is calculated according to $A_y - BN_y$, where A_y stands for the y coordinate of the articulator in question and BN_y for the y coordinate of the bridge of the nose (origin). Whenever two gestures are involved (for instance, any lip aperture gestures), the distance is calculated as the distance between their respective vertical coordinates.

Note that two distinct primitives are used to indicate the articulator entering and leaving the current quantisation region (cluster). For example, if we consider the medium range of the tongue dorsum heights, when the new cluster centroid represents a higher range, we represent this transition by the TD-RaiseMid gesture. Otherwise, if the new cluster centroid represents the lower range, the transition is represented by a

different gesture TD-LowerMid.

5.2 Evaluation Strategy

The evaluation was applied to all 920 gestural structures automatically derived from the utterances of the corpus. Overall, 9,879 phonetic labels were available for the 14 classes corresponding to the 14 ETS supertransforms. In general, a supertransform (phoneme class), was considered to match if any of its constituent transforms matched the gestural structure corresponding to the label.

Since the representation is asynchronous and the articulation of stop consonants is *anticipatory* [Ladefoged, 2001], primitive gestures are not constrained to appear within the phoneme boundaries of any given label. For such anticipatory articulation, the primitive gestures forming constrictions usually appear before the beginning of the phonetic label, often spanning multiple phoneme boundaries. For instance, most of the gestures participating in the articulation of the voiced velar stop $/g/$ shown in Figure 3 appear before the beginning of the corresponding phonetic label. The gesture VC-Touch completing the constriction occurs at 0.266 sec, 94 ms before the phoneme boundary.

Given the above, the search boundaries for any given constituent transform are not restricted to the boundaries of the phonetic label, but also include the boundaries of several previous phonemes. Phonetic boundaries are specified in terms of the start and end times of a particular label. In particular, for each phonetic label and a candidate class element (transform) to be matched, the sought structure is declared as a successful match if it is identified by the search algorithm presented in Section 4 (starting from the end time of the phoneme label and proceeding backward in time) and if one of the following conditions is satisfied: (1) the candidate class element is located within the phoneme boundaries of the phonetic label; (2) it is found to be overlapping with the beginning of a phoneme label (i.e. the formation of the constriction is anticipatory, beginning before the start of a phoneme boundary).

5.3 Results and Discussion

The overall results of the verification of the 14 classes of consonantal phonemes are presented in Table 3 in the form of a confusion matrix. For each of the classes, the number of correct matches is shown on the diagonal in bold. The number of class phonemes which failed to classify in any of the available classes is shown under $/X/$. The number of expected phonemes is given by N_e, while N_o stands for the number of correctly matched phonemes. The accuracy of the structural matching, denoted C, is given in the last column. As can be seen from Table 3, out of 9,878 phonemes, 7,679 were classified correctly and 278 failed to match against any of the available classes. The overall accuracy is 77.74%.

Analysis of the per-class statistics shows that the lowest accuracy of 62.16% was obtained for the alveolar nasal $/n/$ which was often confused with voiced alveolar stop $/d/$. This could be explained by the fact that the postulated class structures of these sounds (see Table 2), are not sufficiently discriminative, differing by only one gesture (the production of $/n/$ is achieved in the presence of the velic opening). Therefore, due to a failure of the pre-processor to detect the corresponding change in the state of the velum, $/n/$ is often classified as $/d/$. The relatively frequent misclassification of the class $/m/$ as $/b/$ can also be attributed to the same cause. In general, it is expected that performance should improve with a more accurate pre-processor and better discriminating phonemic class descriptions, especially in the obvious cases when misclassification is not due to noisy data or to errors in linguistic labelling of the corpus.

6 Summary and Conclusion

We presented a novel structural representation of speech, developed within the Evolving Transformation System formalism. The representational unit chosen was the *gesture*, which is seen as the interaction of the various physiological organs involved in the act of speech production. We described the 14 classes of English consonantal phonemes in terms of non-trivial combinations of articulatory gestures. A structural matching algorithm capable of detecting an instance of one of the above classes inside a gestural structure (corresponding to a speech utterance) was also presented. The performance of the proposed class descriptions on real data (the MOCHA corpus) was evaluated, yielding an overall matching accuracy of 77.74%. Our results support the hypothesis that a structural representation of articulatory speech allows adequate identification of the phonemic classes.

Despite a general agreement that the use of articulatory information is highly beneficial (on both linguistic and physiological grounds), progress in that direction has been limited. This state of affairs may be attributed to a poor understanding of how the various parts of a speech production system interact. We believe that the use of a representational formalism that supports the description of *structural* classes of articulatory processes (such as ETS) will benefit both speech science and the linguistic community. In particular, such a formalism will guide the modelling of these complex speech production mechanisms.

Acknowledgments

The authors would like to thank Lev Goldfarb and the anonymous reviewers for many useful suggestions.

References

[Abeles *et al.*, 2004] M. Abeles, G. Hayon, and D. Lehmann. Modeling Compositionality by Dynamic Binding of Synfire Chains. *Journal of Computational Neuroscience*, 17(2):179–201, 2004.

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

[Forbus, 1996] K. Forbus. Qualitative reasoning. In *CRC Hand-book of Computer Science and Engineering*. CRC Press, 1996. Draft chapter.

[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.

Classes	/b/	/p/	/g/	/k/	/d/	/t/	/v/	/f/	/ng/	/m/	/n/	/ch/	/zh/	/sh/	/X/	N_o/N_e	C (%)
/b/	**495**	0	2	0	1	0	4	81	0	13	0	14	0	0	2	495/612	80.89
/p/	33	**644**	2	0	1	0	0	41	0	0	0	18	0	0	1	644/740	87.03
/g/	6	0	**287**	9	2	0	3	17	6	1	0	16	2	7	28	287/384	72.14
/k/	11	0	22	**794**	13	0	2	33	12	1	0	77	5	27	73	794/1070	74.21
/d/	49	0	10	0	**821**	0	8	20	12	7	0	71	24	13	27	821/1062	77.31
/t/	66	0	31	4	64	**1401**	6	59	5	5	0	0	11	27	63	1401/1742	80.42
/v/	1	0	1	0	1	0	**431**	9	0	0	0	4	0	0	5	431/452	95.35
/f/	11	0	1	0	0	0	5	**503**	0	0	0	0	1	3	2	503/526	95.63
/ng/	11	0	21	0	15	0	5	1	**207**	0	0	0	1	0	19	207/280	73.93
/m/	176	0	3	0	8	0	25	3	0	**600**	0	0	0	0	5	600/820	73.17
/n/	142	0	32	0	277	0	23	0	17	17	**1038**	23	49	1	51	1038/1670	62.16
/ch/	0	0	5	1	0	0	0	15	3	0	0	**156**	1	9	0	156/194	80.41
/zh/	0	0	0	0	0	0	0	0	0	0	0	0	**34**	0	0	34/34	100.00
/sh/	0	0	0	0	1	0	0	0	0	0	0	1	0	**288**	2	288/292	98.63
All															278	7679/9878	77.74

Table 3: Evaluation results for `fsew` and `msak` speakers, shown as a confusion matrix.

[Goldfarb, 2004] L. Goldfarb. Representational formalisms: Why we haven't had one. In L. Goldfarb, editor, *Pattern Representation and the Future of Pattern Recognition (Proc. Satellite Workshop of 17th International Conference on Pattern Recognition)*, pages 3–22, Cambridge, UK, August 2004.

[Gutkin and King, 2005] Alexander Gutkin and Simon King. Detection of Symbolic Gestural Events in Articulatory Data for Use in Structural Representations of Continuous Speech. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume I, pages 885–888, Philadelphia, PA, March 2005.

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

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

[Jung *et al.*, 1996] Tzyy-Ping Jung, A. K. Krishnamurthy, S. C. Ahalt, M. E. Beckman, and Sook-Hyang Lee. Deriving gestural scores from articulatory-movement records using weighted temporal decomposition. *IEEE Trans. Speech and Audio Processing*, 4(1):2–18, January 1996.

[Kanungo *et al.*, 2002] T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu. An Efficient k-Means Clustering Algorithm: Analysis and Implementation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(7):881–892, 2002.

[Kaplan, 1971] Harold M. Kaplan. *Anatomy and Physiology of Speech*. McGraw-Hill, 2nd edition, 1971.

[Kennedy and Goldstein, 2003] M. Studdert Kennedy and L. M. Goldstein. Launching language: The gestural origin of discrete infinity. In Morten H. Christiansen and Simon Kirby, editors, *Language Evolution*, Studies in the Evolution of Language. OUP, New York, 2003.

[Ladefoged, 2001] P. Ladefoged. *A Course in Phonetics*. Harcourt Brace Jovanovich, 4th edition, 2001.

[Nguyen, 2000] N. Nguyen. A Matlab toolbox for the analysis of articulatory data in the production of speech. *Behaviour Research Methods, Instruments and Computers*, 32:464–467, 2000.

[Pavlidis, 2003] T. Pavlidis. 36 years on the pattern recognition front. *Pattern Recognition Letters*, 24:1–7, 2003.

[Talkin, 1995] D. Talkin. Robust algorithm for pitch tracking. In W. B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Synthesis*. Elsevier Science B.V., 1995.

[Valiente, 2002] Gabriel Valiente. *Algorithms on Trees and Graphs*. Springer-Verlag, Berlin, 2002.

[Wrench, 2000] A. A. Wrench. A multichannel articulatory database for continuous speech recognition research. In *Phonus5, Proc. Workshop on Phonetics and Phonology in ASR*, pages 1–13, University of Saarland, 2000.

[Zemlin, 1968] Willard R. Zemlin. *Speech and Hearing Science: Anatomy and Physiology*. Prentice-Hall, New Jersey, 1968.