On the Articulatory Representation of Speech within the Evolving Transformation System Formalism

Alexander Gutkin[1], David Gay[2], Lev Goldfarb[2], and Mirjam Wester[1]

[1] Centre for Speech Technology Research, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW, United Kingdom
alexander.gutkin@ed.ac.uk, mwester@inf.ed.ac.uk
[2] Faculty of Computer Science, University of New Brunswick
540 Windsor Street, P.O. Box 4400, Fredericton, NB, Canada
goldfarb@unb.ca, dave.gay@unb.ca

Abstract. This paper deals with the formulation of an alternative, structural, approach to the speech representation and recognition problem. In this approach, we require both the representation and the learning algorithms to be linguistically meaningful and to naturally represent the linguistic data at hand. This allows the speech recognition system to discover the emergent combinatorial structure of the linguistic classes. The proposed approach is developed within the ETS formalism, the first formalism in applied mathematics specifically designed to address the issues of class and object/event representation. We present an initial application of ETS to the articulatory modelling of speech based on elementary physiological gestures that can be reliably represented as the ETS primitives. We discuss the advantages of this gestural approach over prevalent methods and its promising potential to mathematical modelling and representation in linguistics.

1 Introduction

Human speech is first and foremost a communicative signal, its purpose being to convey meaning [10]. While production and perception of speech can be easily accomplished by humans, the respective modelling of speech by machines is a complicated task. Despite significant progress made over the last twenty years [26], there is still a considerable amount of work that needs to be done before the results can be declared satisfactory. One of the main reasons for this situation in automatic speech recognition research, in our view, is due to the apparent lack of suitable structural formalisms that can support the complex linguistic representations (this state of affairs is not restricted to speech recognition and appears to apply to the pattern recognition field in general [19, Section 3]).

In traditional phonology, speech is represented by linearly concatenating phonemes [6] which are then mapped to the physical units of sound via an allophonic level. The allophonic level models the co-articulatory and prosodic variations with the help of allophonic rules [16]. Concatenation of the outputs of the allophonic rules results in the systematic transcription of speech. Modern research in speech production and perception has shown that this view of speech as consisting of concatenated invariant static units, disagrees with reality. One of the modern phonological theories that tries to address the failure of finding satisfactory fundamental units of analysis is the theory of articulatory phonology proposed by Browman and Goldstein [4]. In articulatory phonology, instead of looking at a shallow

A. Gutkin, D. Gay, L. Goldfarb and M. Wester: On the Articulatory Representation of Speech within the Evolving Transformation System Formalism: In Proc. Workshop on Pattern Representation and the Future of Pattern Recognition (17th International Conference on Pattern Recognition), (L. Goldfarb, ed.) Cambridge, UK (2004) pp. 57–76

description of the act of speech production using traditional units, such as phonemes represented as bundles of phonological features [11], the vocal tract action during the speech production is decomposed into discrete re-combinable atomic units. The central idea is that while the observed resulting products of articulation (articulatory and acoustic measurements) are continuous and context-dependent, the actions which engage the organs of the vocal tract and regulate the motion of the articulators are discrete and context-independent [3]. 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 [9, 22], thus sharing these combinatorial, *self-diversifying* [1] properties with other natural systems, known from chemistry and developmental biology.

In this paper, an initial application of the Evolving Transformation System (ETS) formalism [7, 8] to the domain of spoken language is described which has been inspired by the physiological combinatorial outlook on speech advocated by the theory of articulatory phonology. The aim of this work is not to rigorously put the articulatory phonological theory of [4] in the ETS setting, but rather to show how some of the most fundamental ideas of the articulatory physiological approach to speech, such as the combinatorial structure hypothesis, can be formally specified within the ETS formalism.

This paper consists of four sections. The main tenet of articulatory representation, namely the relation of physiological articulatory gestures to the abstract units of information in the ETS formalism, called *primitives* is introduced in Section 2. Section 3 describes an initial articulatory-inspired ETS representation for a limited class of consonantal phonemes, focusing on elaborating the choice of abstract gestural units upon which the representation rests, describing the major stages of the design and reporting the results of initial experiments. A brief overview of the existing approaches to speech modelling, their comparison to the ETS representation, followed by discussion of the potential benefits of introducing scientific representational formalisms into the linguistic and speech recognition communities is presented in Section 4. Conclusions and an outline of the future work that we intend to undertake is given in Section 5.

2 Physiological Gestures as ETS Primitives

In this section, we describe the main working assumptions for the articulatory representation of speech. In the ETS formalism, the concept of *primitive events* is one of the most fundamental [7]. Primitives are the basic atomic building blocks of the model which combine together to form complex structures (events) and transformations. In a hierarchical representation, each next level primitive (except for the ones at the initial, sensory level) can be expressed as a complex structure, in turn consisting of the current level primitives forming complex structures called *class supertransforms*.

In line with the process, event-based, philosophy of the ETS formalism [7], we base our analysis on the various articulatory processes (gestural events and combinations thereof) which operate and cause changes in the states of the articulatory organs. Thus, the "objects" under investigation become the various organs involved in the production of speech, while the dynamic interaction between these organs (which phoneticians call *processes*, e.g. a nasal sound is a result of an "oro-nasal process"), resulting in speech, is described by the ETS primitive events. The choice of initial level ETS primitives, therefore amounts to first identifying the articulators participating in speech production based on physiological [15] and phonetic [16] evidence (the articulators are chosen to correspond to the *sites* of the ETS primitives), and second, selecting the most distinct gestures (corresponding to *names* of the ETS primitives) involving the articulators specified above. The theoretical motivation for this more general outlook on the articulators and the interactions between them is supported by linguistic theory, which states that an analysis on a lower, motor level introduces too much anatomical detail which is linguistically irrelevant for the discrimination between various sound patterns [16].

Although the articulators share some mechanical degrees of freedom, they are commonly assumed to be anatomically distinct and independent (any constriction formed by one of the organs does not necessarily produce a constriction in any other) [9]. This is reflected in the choice of the sites of ETS primitives within this representation which are all different (there are no sites of the same type).

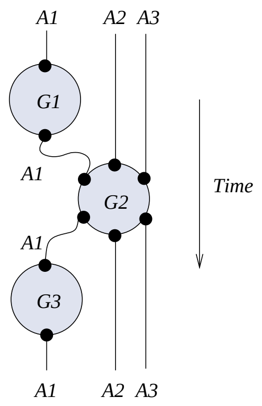

Fig. 1. Pictorial view of a gestural structure

Figure 1 shows an abstract articulation involving articulatory organs $A1$, $A2$, $A3$ and three gestures $G1$, $G2$ and $G3$ making use of these organs (the vertical positioning of the gestures corresponds to the actual flow in time of the pre-processing algorithm which detects them). The gesture $G1$ operates on one articulator $A1$ only, whereas gesture $G2$ involves all of the depicted articulators and follows $G1$. Gesture $G1$ might mean "raise $A1$", gesture $G2$ might mean "move $A1$ to $A2$ while $A3$ vibrates", while gesture $G3$ could mean "lower $A1$". Within the ETS formalism, this pictorial representation corresponds to the temporal sequence of three *primitives* $G1[A1|A1]$, $G2[A1, A2, A3|A1, A2, A3]$ and $G3[A1|A1]$ which form *struct* $G_\sigma = [G1 \dashv G2 \dashv G3]$ representing some non-trivial gesture.

It is not difficult to see, that each primitive gesture, represented by an ETS primitive, encapsulates both syntactic and semantic information. The syntactic information, allows for structural processing by the appropriate training and recognition algorithms defined within the ETS framework [7], while the semantic information makes the representation meaningful and fully interpretable.

3 Gestural Representation of Speech: An Example

In this section, a tentative application of ETS to the speech representation domain is presented. For the sake of clarity and simplicity, we have chosen to concentrate on a rather limited set of phonemes: the two pairs of velar ($/k/$ and $/g/$) and bilabial ($/p/$ and $/b/$) stops. In what follows, we first describe the articulatory corpus used for the experiments, then introduce the primitive gestures for the velar and bilabial stops and describe the ways in which the gestures are extracted from continuous speech. In Section 3.3, an evaluation of the reliability of the gestural primitives is reported. This is followed by a description of how the primitive gestures combine to form ETS structs. After explaining the structs, we go into more details concerning some other fundamental ETS concepts: transformations and class supertransformations. This Section 3 concludes with a few thoughts and ideas pertaining to what a full ETS speech recognition system might look like.

3.1 The Articulatory Corpus

The articulatory corpus which we used for the experiments is the MOCHA corpus [24, 25], which is becoming more popular with the automatic speech recognition community (articulatory research has traditionally received more attention from linguists [20]) as more researchers become interested in using articulatory parameters either as a supplement to or substitute for spectrally based input parameters. The MOCHA corpus consists of articulatory and acoustic recordings of 460 phonetically-rich sentences designed to provide good phonetic coverage of English. At the moment, the database contains the finalised recordings for one male and one female speaker, each consisting of approximately 31 minutes of speech. The particular dataset which we used came from the recording of a female speaker of British English (acronym `fsew0`).

The articulatory channels include Electromagnetic Articulograph (EMA) sensors directly attached to the upper and lower lips, lower incisor (jaw), tongue tip (5-10 mm from the tip), tongue blade (approximately 2-3 cm posterior to the tongue tip sensor), tongue back (dorsum) (approximately 2-3 cm posterior to the tongue blade sensor) and soft palate (velum). The EMA data has been recorded at 500 Hz. Coils attached to the bridge of the nose provided the frame of reference.

Laryngograph/EGG measures changes in the contact area of the vocal folds, providing the recording of the laryngeal waveform, from which pitch and voiced/unvoiced information can be derived. Both the laryngeal and acoustic waveforms were recorded at 16 kHz.

Electropalatograph (EPG) measurements provide tongue-palate contact data at 62 normalised positions on the hard palate, defined by landmarks on the upper maxilla [24], augmenting some of the information missing from the EMA data. This is produced by the subject wearing an artificial palate specially moulded to fit their hard palate with the 62 electrodes mounted on the surface to detect lingual contact. Each EPG frame (the EPG.3 version of the device was used), sampled at 200 Hz consists of 96 bits, 34 bits of which are unused. Each bit from the 62 bit mask is on if the contact was detected, off otherwise.

The articulatory data was post-processed to synchronise the channels and correct for the EMA head movement and discrepancies in coil placements during the recording. The resulting coordinate system of EMA trajectories consisting of (x, y) coordinates has its origin at the bridge of the nose, with positive x direction being towards the back of the vocal tract, away from the teeth, and positive y direction being upwards towards the roof of the mouth.

The corpus was automatically labelled using forced alignment of the acoustic signal with phone sequences generated from a phonemic dictionary, thus phonetic labels are available (see [24, 25] for more information).

3.2 Emerging Gestural Primitives

Following the guideline outlined in Section 2, the critical organs participating in the articulation of velar and bilabial stops are identified first. The articulation of velar closures can be characterised by the trajectory of the tongue dorsum (which is usually high during articulation) and velum (which is touched by the tongue dorsum to achieve closure). Velar closures usually have a short duration and are promptly released to prepare for the next articulation. The bilabial closures are characterised by the trajectories of both upper and lower lips, constriction being achieved by lip closure, followed by the release of the lips resulting in the release of air from the oral cavity. In addition, bilabial and velar closures can be either voiced (vibrating vocal folds) or unvoiced (with no vibration in the vocal folds). The organs thus identified, become the sites of the ETS primitives corresponding to primitive articulatory gestures and are shown in Table 1, together with the corresponding semantics and types of measurements available for these organs in the MOCHA database.

Table 1. Sites of ETS primitives corresponding to primitive gestures

Site	Semantics	Measurement Type
UL	upper lip	EMA
LL	lower lip	EMA
TD	tongue back (dorsum)	EMA, EPG
VL	soft palate (velum)	EMA, EPG
VF	vocal folds	laryngeal, acoustic

The atomic articulatory gestures, modelled as ETS primitives, comprising the critical articulations which distinguish between bilabial and velar stop consonants are derived from the measurements available in the MOCHA database. The resulting primitives are presented in Fig. 2 and can be roughly subdivided into four groups: the articulatory gestures of the vocal folds, resulting in voiced or unvoiced sounds, the two gestures participating in velar closure, the gestures controlling the aperture of the lips (bilabial closure) and the gestures describing the vertical trajectory of the tongue dorsum.

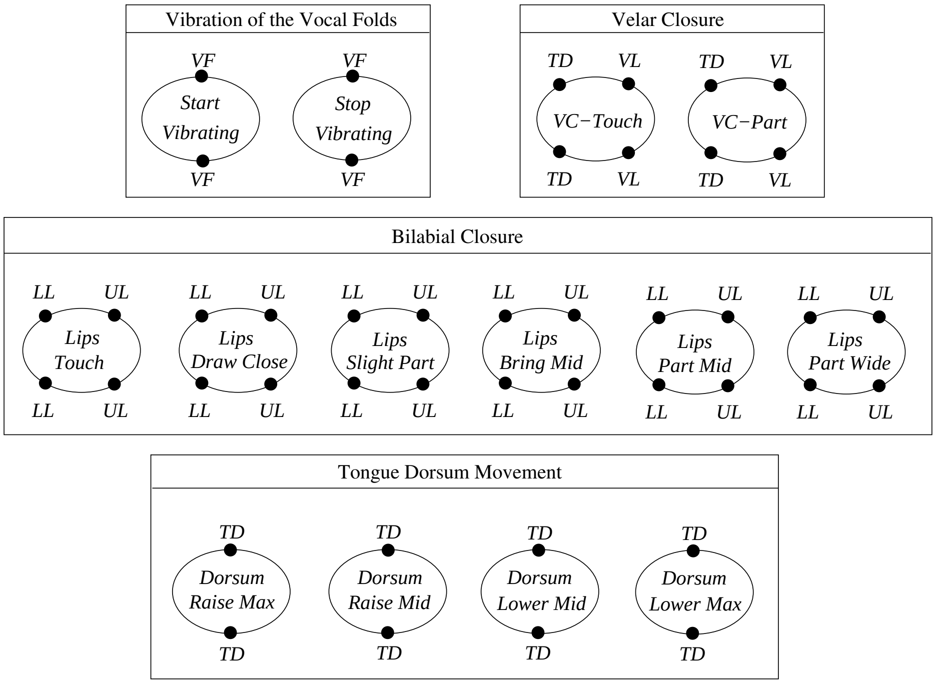

Fig. 2. Initial level ETS primitives subdivided into related groups

Vibration of the vocal folds that defines voiced and unvoiced sound patterns is represented by the two primitives standing for the beginning and end of vibration respectively. We used the pitch detection algorithm described in [23] to extract pitch information from both the laryngeal and acoustic recordings provided in the MOCHA database. We used the 5 ms interval for analysis frames and the pitch frequency search range between 25 Hz and 600 Hz. At any given point in time, the decision whether the vocal folds have started or stopped vibrating 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 frames of a 16 kHz recording), which is the average duration of a typical short vowel.

The algorithm detecting the gestures causing velar closure makes use of the electropalatographic (EPG) data provided by MOCHA. In general, the output of the EPG sensor, at any given point in time, consists of 8 8-bit binary vectors with a simple spatial structure. The first three rows represent the alveolar region (the first and 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. We use the velar contact index measured by the linear combination of the rows representing the velar region (which is a sum of all the bits of the last three rows) and the centre of gravity index (COG), given by

$$COG = \frac{\sum_{i=1}^{8}(8-i+1)R_i}{\sum_{i=1}^{8} R_i} \ ,$$

where R_i denotes the number of contacts in row i [18]. These measurements provide reliable estimates of tongue contact in the velar region. The velar closure primitive *VC-Touch* emerges when the velar index increases beyond the velar threshold of 14 (there are 24 sensors in the velar region) and the value of the COG index is less than 4 (indicating a shift in the centre of gravity to the palato-velar region). The primitive gesture *VC-Part*, specifying the release of the velar closure, appears when the velar index decreases below the velar index threshold of 14 (the COG value is ignored in this case). 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 indices to be steady for any period of time.

The remaining two groups of primitives, the bilabial and tongue dorsum groups, are detected from the EMA trajectories. In general, given the distance measurements between any pair of articulators, we quantise these values. We chose the number of clusters N to be four for the detection of bilabial gestures and three for the tongue dorsum gestures respectively. The quantisation, making use of a straightforward k-means procedure, was applied to the entire data available for the female British speaker. Each cluster centroid 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 10 frames (20 ms for the EMA data sampled at 500 Hz), the decision is made to fire a 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 M times. The formula for calculating the lip aperture is $UL_y - LL_y$, where UL_y and LL_y are the y-coordinates of the upper and lower lip, respectively (lip protrusion expressed by the x-direction is ignored). The tongue dorsum height is calculated by the formula $TD_y - BN_y$, where the TD_y and BN_y stand for the y-coordinates of the tongue dorsum and bridge of the nose (the origin), respectively.

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 $DorsumRaiseMid$ gesture. Otherwise, if the new cluster centroid represents the lower lange, the transition is represented by a different gesture $DorsumLowerMid$.

The gesture specifying the vertical trajectory of the tongue dorsum provides additional information which helps to determine velar closures, since a high position of the tongue dorsum associated with the $DorsumRaiseMax$ gesture, is a prerequisite for the formation of the velar constriction.

3.3 Evaluation of Gestural Primitives

In order to evaluate the reliability of the primitive gestures described above, experiments were conducted to assess the potential accuracy of their detection. The evaluation was conducted on the entire data set available for the female British English speaker (acronym `fsew0`), for which 460 utterances with an overall duration of approximately 31 minutes are available. Since the corpus provides the phonetic labels (obtained by an automatic alignment procedure [24]), it is possible to check whether any of the primitive gestures which are *a priori* known to participate in the production of the four phonemes /*b*/, /*p*/, /*g*/ and /*k*/ - actually appear during runtime. This *a priori* knowledge is derived from phonetics [16]. Table 2 shows the list of phonemes (bilabial and velar stops) we used in the experiments. For each phone, the frequency of occurrence of the corresponding label in the corpus is shown, along with the list of primitive gestures which are *a priori* hypothesised to participate in the formation of that phone. Note that for any given phone, all the required gestures belong to three different groups (consisting of bilabial, velar and vocal folds gestures). For example, according to the information provided by the phonetic labels, there are 192 instances of the unvoiced bilabial stop /*p*/ in the corpus for the `fsew0` speaker. The gestures which specify the formation of the unvoiced bilabial stop are: the vibration of the vocal folds stopping (*VF StopVibrating*) and the closure of the lip aperture (*LipsTouch*). In addition, the lack of constriction between the tongue dorsum and the velum is specified as a necessary prerequisite (*VC-Part*), since the formation of the bilabial closure in English cannot be accomplished with the velar constriction being held at the same time.

Table 2. Primitive gestures hypothesised to form bilabial and velar articulatory constrictions

Phoneme	Frequency	Names of primitive gestures		
/*b*/	306	*VF StartVibrating*	*LipsTouch*	*VC-Part*
/*p*/	192	*VF StopVibrating*	*LipsTouch*	*VC-Part*
/*g*/	535	*VF StartVibrating*	*VC-Touch*	*DorsumRaiseMax*
/*k*/	370	*VF StopVibrating*	*VC-Touch*	*DorsumRaiseMax*

We have chosen to test the gestures participating in the formation of the constrictions rather than releases of the constrictions. For example, during the hypothesised articulation of /*k*/, we test for the velar closure caused by tongue dorsum touching the velum.

The verification algorithm is applied to all the utterances in the corpus. For each phonetic label from a given utterance, each of the primitive gestures from a corresponding list is processed in turn. According to the algorithm, the primitive gesture *participates* in the formation of the corresponding phone if one of the following conditions is satisfied:

– The primitive gesture appears within the boundaries (specified by the start and end times) of the phone label currently being processed.

– The primitive gesture occurs somewhere within the boundaries of several previous phones. In this case, the algorithm checks that no other primitive gesture belonging to the same group occurred between the current phone and the phone where the primitive gesture of interest was detected. This is to ensure that the primitive gesture being verified (for example, *LipsTouch*) is not later cancelled by some other primitive gesture from the same group (for example, *LipsSlightPart*) before the current phone boundaries.

The evaluation experiment was conducted on the entire `fsew0` dataset which has 1403 phonetic labels for bilabial and velar stop consonants and the results are presented in Table 3. For each primitive gesture, the expected frequency of occurrence, the actual observed frequency and the percentage of error (representing the gestures which did not occur where expected) are shown. As can be seen, the least reliable gesture is the start of vocal folds vibration, exhibiting an error of 8.43%, which is due to the fact that the pitch tracker was not specifically tuned for this particular female speaker (the pitch search range is too wide) and the algorithm's default parameters were used. This will be improved upon in the future. Overall, out of 4209 expected gestures, 4076 appear to have been detected correctly, with a reasonably low error of 3.16%.

Table 3. Evaluation results for each of the primitive gestures

Name of the Gesture	Observed Frequency	Expected Frequency	Error (%)
LipsTouch	673	676	0.44
VC-Touch	675	727	7.15
VC-Part	670	676	0.89
DorsumRaiseMax	715	727	1.65
VF StartVibrating	456	498	8.43
VF StopVibrating	887	905	1.99
Total	4076	4209	3.16

Since the corpus was autolabelled, it contains transcription errors due to pronunciation variants not catered for in the single pronunciation dictionary, reading errors and co-articulation processes (which might result in deletion or alteration of a sound depending on its phonetic context) [24]. The artifacts of the autolabelling may account for a percentage of the errors encountered during the evaluation. Careful examination of the autolabelling process is needed for a better understanding of the problematic transcriptions. Improving the quality of the transcriptions will most likely result in the overall accuracy improvements.

3.4 Temporal Sequences of Primitive Gestures as ETS Structs

Temporal sequences of primitive gestures form *gestural structures*, or simply *structs* in the language of ETS formalism. Figure 3 shows the gestural structure, represented as an ETS struct, detected in the MOCHA database. This gestural structure corresponds to the word "coconut" (transcribed using the CMU labels as /*k ou k @ n uh t*/). The time of the detection for each of the primitives is shown in seconds along with the phonetic labels provided for that fragment. The first label /*sil*/ marks the beginning of an utterance.

Turning to the analysis of the unvoiced velar stop expressed by the phoneme /*k*/, the corresponding struct in Fig. 3 has two instances of it. The sequence of primitive gestures resulting in a first velar constriction starts with the tongue dorsum reaching its maximum height at about 0.400 sec, followed by the contact between the tongue dorsum and the velum at 0.466 sec. The constriction is completed by 0.484 sec at which point the vocal folds stop vibrating resulting in a complete unvoiced velar closure. The constriction is released at 0.596 sec, at which point the tongue dorsum and the velum part. This is followed by the lowering of the tongue dorsum. The second instance of a velar stop and release begins with the tongue dorsum reaching its maximum height at 0.614 sec followed by velar contact at 0.686 sec, with vocal folds ceasing to vibrate at 0.718 sec. The release sequence starts with the vocal folds vibration at 0.764 sec, followed by the lowering of the tongue dorsum. Note that the articulatory gestures participating in the formation of bilabial sounds occur independently of the formation of velar sounds, as can be explicitly seen in the corresponding struct. The above analysis is made possible by the fact that the gestural structure thus represented carries within itself both the syntactic and semantic information expressed by the ETS primitives standing for primitive gestures. Syntactic information provides the means of enforcing the allowable structure on the gestural combinations, while the semantic information is expressed by both the structure of each single primitive and the overall gestural structure.

It is instructive to see that the gestural structure of Fig. 3 exhibits asynchrony with respect to the phone label boundaries. For instance, the constriction corresponding to the second instance of an unvoiced velar stop completes 24 ms before the beginning of /*k*/ (at the end of /*ou*/) and is held for 32 ms with the release starting 8 ms after the boundary of /*k*/. This further supports the hypothesis that phones are not the correct unit for speech analysis as they make it near impossible to account for the asynchrony and dynamic nature of the speech production process. However, since at present, no articulatory corpus labelled at a gestural level is known to exist, phonetic (or syllabic) labels are the only means of veryfing the reliability of the pre-processing and recognition stages.

3.5 Common Recurring Gestural Patterns as ETS Transformations

By examining the ETS gestural structs, generated by the preprocessing algorithm described previously, several structurally and semantically related gestural fragments of the structs can be discerned. For each of the sound patterns under investigation, namely velar and bilabial stop consonants, the corresponding gestural fragments can be roughly divided into two parts, the actual constriction and the release. As mentioned in the previous section, the primitives comprising the two parts of the corresponding gestural fragment exhibit asynchrony and often span multiple phone boundaries (the anticipatory movement toward

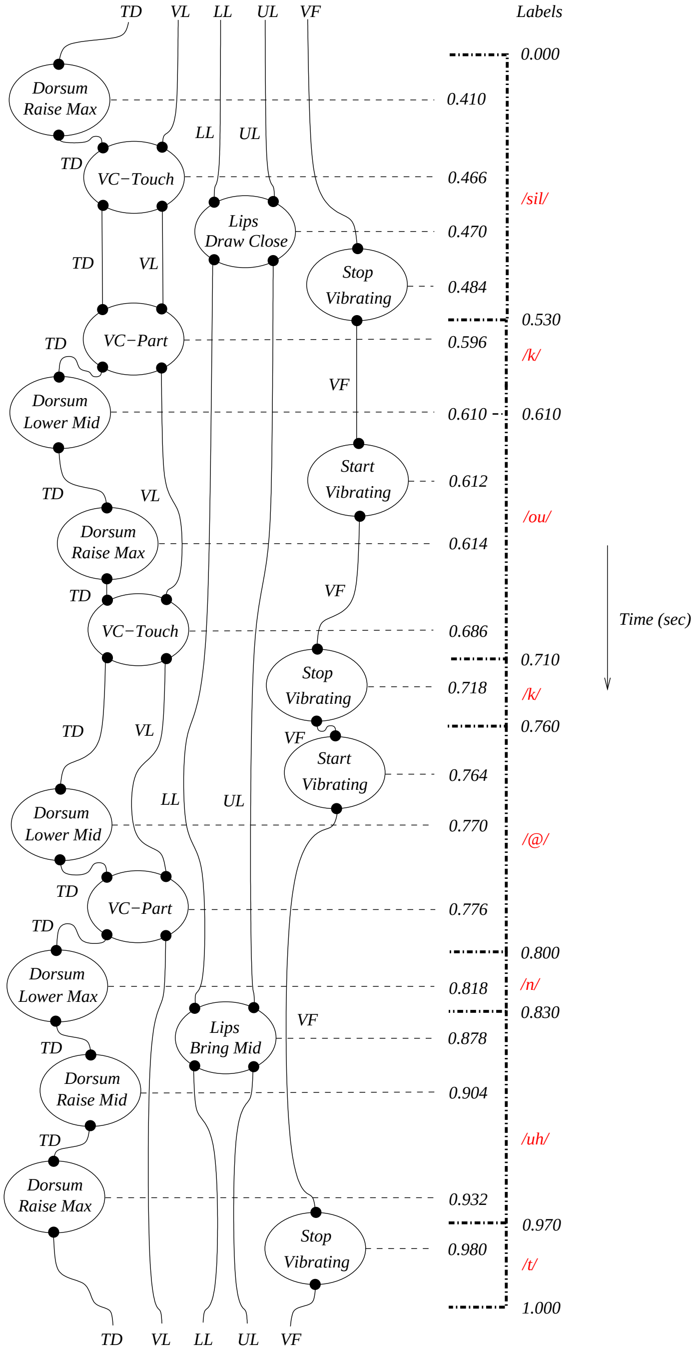

Fig. 3. ETS struct representing the word "coconut" (transcribed as $/k\ ou\ k\ @\ n\ uh\ t/$). Time of the appearance (in seconds) of each of the primitive gestures is shown along with the available phonetic labels

the lip constriction target, for instance, might start relatively early, before the constriction is actually satisfied). Therefore, in our analysis, the constriction starts with the first primitive gesture aimed at producing this constriction, ending with the last primitive gesture which secures its release.

Figure 4 shows some of the common gestural patterns, four per sound, encountered in the data for unvoiced velar (top) and bilabial (bottom) stops (the fragments for their voiced counterparts are not shown). The body of each of the ETS transformations [7] consists of the sequence of gestures which participate in the release of the stop, while the gestures which participate in the formation of the actual constriction are depicted as part of the transformation context. The context of the transformation can thus be seen as a necessary precondition for the respective sound to be produced (the gestures which are not critical for a particular articulation are shown with the connections to them crossed out).

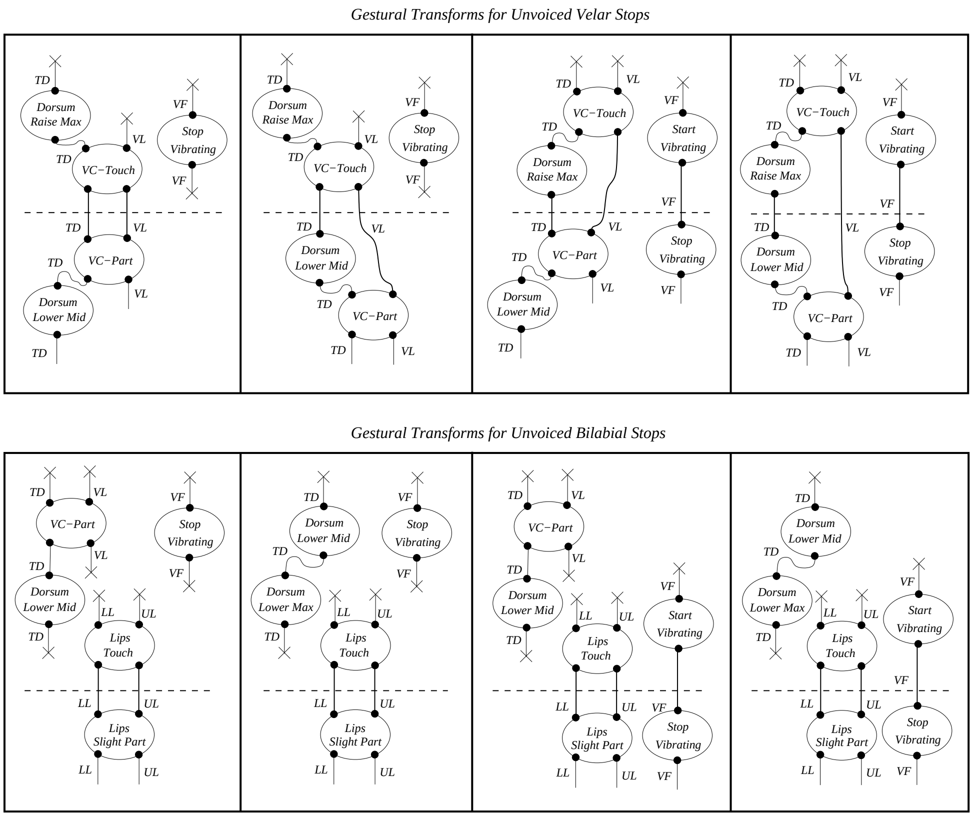

Fig. 4. Some common gestural patterns encountered in the data for unvoiced velar (top) and bilabial (bottom) stops represented as ETS transformations (their voiced counterparts as well as the times of occurrence of the constituent primitive gestures are not shown)

Note that while each of the transformations has a similar higher-level semantics (for instance, all four transformations shown in the top figure represent the release of a velar stop), structurally they are all different. The first two transformations differ in their bodies which can be interpreted as follows: For the first body, the release is accomplished by first removing the tongue back from the velum (the position of the tongue in the oral cavity is still high) followed by the lowering of the tongue. For the body of a second transformation, the tongue appears to be lowered (together with velum) and only then detached. For all of the transformations, the vocal folds may have already stopped vibrating, meeting the necessary but not sufficient requirement for the articulation (in which case they are shown in the contexts) or, they stop vibrating at the onset of the release of that particular stop (in which case they are in the bodies of the transformations).

3.6 Phonemic Class Representation via ETS Supertransformations

Here we briefly describe the ongoing work on the class representation of phonemes via the ETS supertransformations and provide a preliminary outline of the recognition process.

An ensemble of semantically and structurally related gestural transformations, described previously, can be represented in the ETS formalism using the concept of a *supertransform* defined in [7], which is a generalisation of the concept of a transformation, introduced earlier. An example of a supertransform for an unvoiced velar stop, consisting of four constituent transformations frequently seen in the MOCHA data, shown in Fig. 5. Formally, the constituent transforms, which are discovered during the training stage, have weights, which are not shown. Each row of the pictorial representation of a supertransform, constructed during the training stage, consists of gestural transformations which share the same contexts (with each context providing the structural means of describing the constriction). Each column of the supertransform representation consists of transformations which have structurally identical bodies (with each body specifying the constriction release). The thicker lines connecting constituent gestures between a body and a context of each constituent transform denote the interface sites which are used to indicate that the articulatory precondition (provided by the context) for the appearance of the anticipated gestural structure (provided by the body) has been met.

The concept of a supertransform is the central one within the ETS formalism since it encapsulates the means of class description. Ensembles of related gestural transformations, like the simple gestural supertransform shown in Fig. 5, provide the means of describing classes of non-trivial gestural events producing similar sounds. The particular supertransform shown here, represents the class of sounds which in linguistics would be labelled using the phoneme /*k*/ and broadly could be described as a sound characterised by multiple possible ways (observed in the training data) of forming and releasing an unvoiced velar constriction. Since the constituent gestures of each supertransform have weights that are learned during training, some of the constituent gestures would be more likely than others (with the less likely ones either being very rare or being the artifacts of noise in the data).

The class of sounds thus defined by a corresponding supertransform becomes the next-level (non-trivial) gesture in the representational hierarchy. In particular, the bodies of the constituent transform leading to this non-trivial primitive specify the various instances of

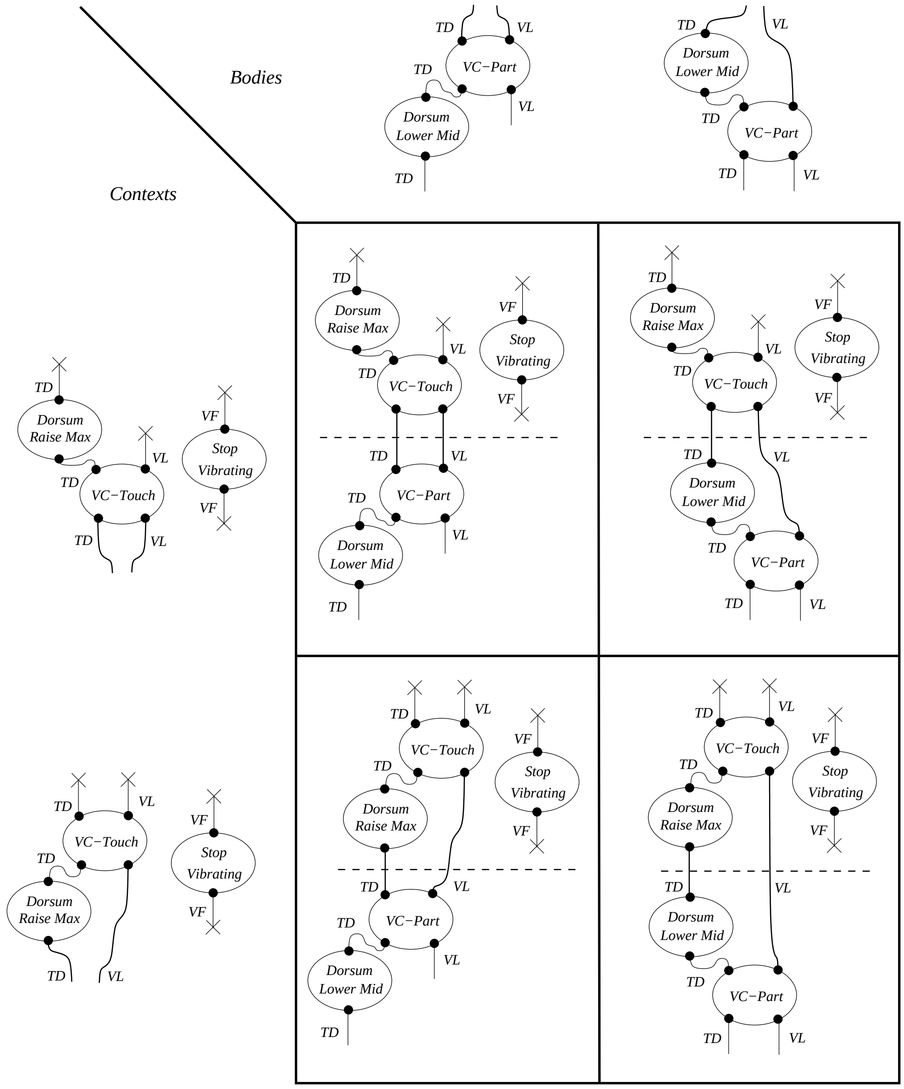

Fig. 5. Example of a class representation of an unvoiced velar stop depicted as an ETS supertransform

this non-trivial gesture observed in the data. Figure 6 shows the emergence of the next-level non-trivial gestures (shown as the next-level ETS primitives on the right-hand side of the figure). The unvoiced velar stop consonants from the initial-level gestural struct from Fig. 3 are shown on the left-hand side. The shading shown indicates two different instances of initial-level gestural events which are represented at the next-level as the ETS primitive $/k/$. Each of those instances corresponds to a different constituent transform from a class supertransform for $/k/$ which is shown in the center. The sites of a next level primitive represent the organs participating in the formation of the class of events it represents. In the case of $/k/$, the sites stand for the tongue dorsum, the velum and the vocal folds.

In very general terms, the recognition of any given phoneme, like $/k/$, therefore reduces to a search in the incoming gestural struct (produced by a pre-processing front-end) for the initial-level gestural transforms structurally similar to any of the constituent transforms from a class supertransform representing a phoneme. This process is aided by the ETS distinguishing between the concepts of context and body within the transformation.

4 Combinatorial Structure of Speech and ETS

As was mentioned earlier, within the ETS formalism, speech can be viewed as relying on a finite (and relatively small) set of underlying language-specific atomic units, or primitive gestures, an assumption which appears to be in agreement with articulatory phonology theory [4]. The atomic gestures asynchronously combine together (subject to anatomical and phonological restrictions) in various ways to form higher-level complex gestural structures (an appropriate analogy would be with the non-trivial compounds in chemistry and other natural sciences [1]) giving rise to various linguistic units, described by phones, syllables, words and so on. In particular, it becomes possible to talk about the *structure* of the distinctive phonological features [11, 6] formed by the primitive gestures in a natural way. With the appropriate choice of the primitive gestures, this can be seamlessly accomplished within an ETS framework.

The primitive gestures, represented by the ETS primitives (see Section 2 for a rationale) are the backbone of this representation and are of primary importance for the understanding of speech production and perception, linking the phonological scientific hypothesis (the combinatorial outlook on spoken language [3, 4, 9, 22]) and the perceived reality. This particular choice of physiological gestures as abstract information units (ETS primitives integrating syntactic and semantic information) for the modelling, has motivated the choice of the physiological corpora (the MOCHA database) for the experiments, since this appears to be the most natural way to proceed (an assumption reinforced by the results of the studies described in [13, 14]). This approach of working directly with data that represents the actual recording of a gestural structure can definitely be preferred over the derivation of the gestural units from acoustic recordings, the popularity of which is driven by the need for having the convenient technological means of recording, by performing some non-linear and scientifically ad-hoc mapping from the acoustics to the physiology. However, obtaining physiological corpora will always remain an issue, whereas acoustic recordings are rife.

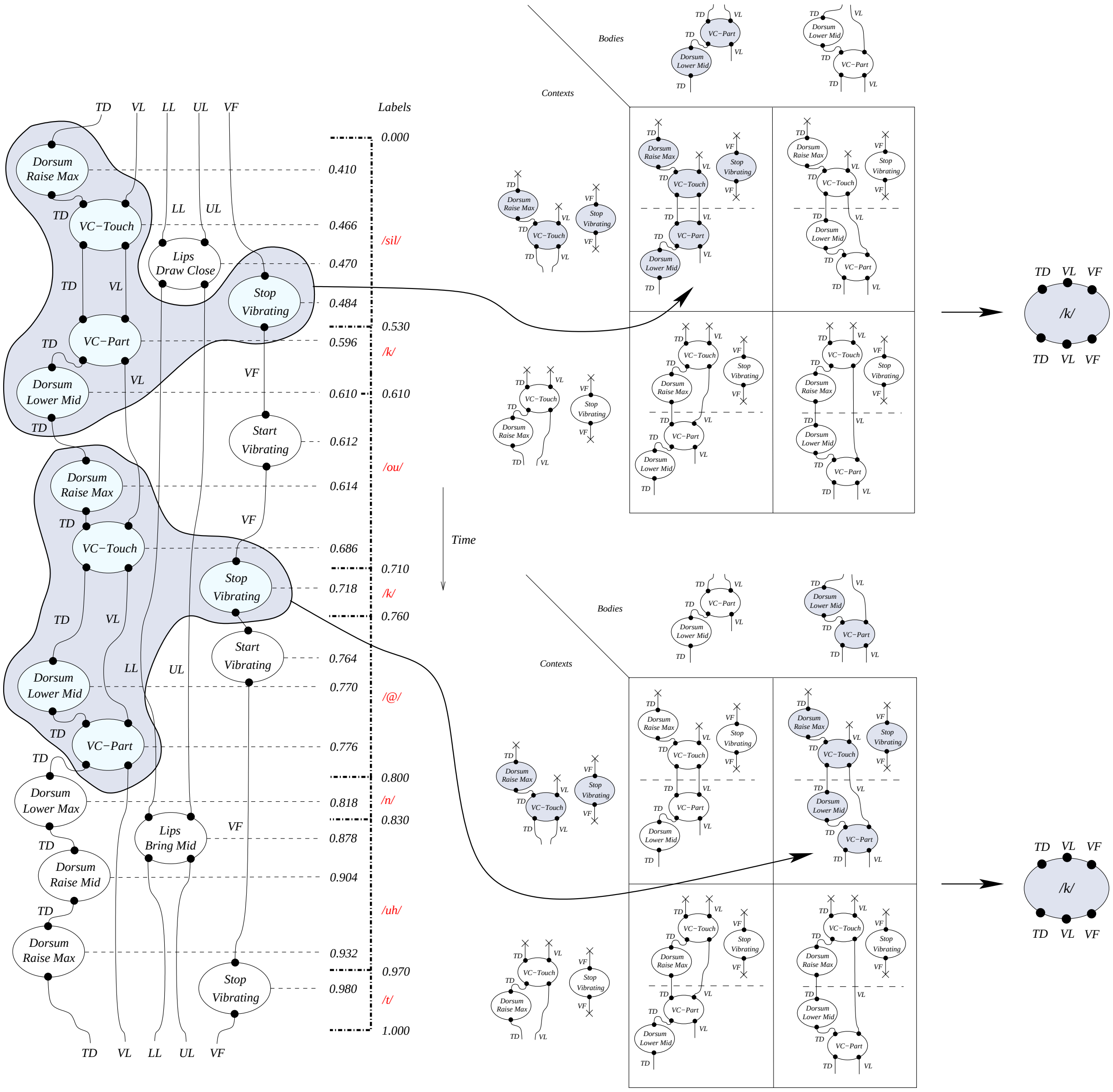

Fig. 6. An outline of the phone recognition process via the ETS supertransforms. The emergence of the two instances of the next-level primitive /k/ (via the class supertransform) from the initial-level gestural struct is shown.

Among the multitude of models which are used when working with speech, statistical approaches to speech recognition are by far the most popular [26]. In a statistical setting, however, the problem of recognition reduces to a risk minimisation problem with a certain set of constraints, rather than the discovery of the (class) structure of the phenomena in question. Despite the fact that the motivation for the choice of basic acoustic units representing speech is often dictated by linguistic considerations [21], we cannot consider these abstract models relying on the corresponding stochastic processes as a representation useful for purposes of further linguistic interpretation.

This is not surprising, since, in words of Jelinek [12, p. 10]: "These models have no more than a mathematical reality. No claims whatever can conceivably be made about their relation to human's actual speech production or recognition". This drawback of the statistical models makes them unsuitable for use as representations in either an acoustic or articulatory setting. However, this should not be viewed as a criticism of this particular approach since it has been specifically developed for a different task without having in mind any of the issues motivating the *structural* representation. As a result, statistical models have no reliable means of representing gestural structure.

Another popular paradigm, used primarily for speech representation, are linguistic annotation graphs, which cover any descriptive or analytic notation applied to raw speech and language data. The notations might come from a wide spectrum of sources ranging from phonological features to discourse structures, morphological and syntactic analysis, word senses, semantic relations and so on, and are usually produced by human labellers. Among the multitude of available multilayer graph-based formalisms, we can discern the single common most important feature: the ability to associate a label or an ordered set of labels with a stretch of time in the recorded speech [2]. An obvious advantage of such a formalism is that with schemes like Annotation Graphs [2], one has a single multilayer graph structure (in which various knowledge sources are unified) associated with a given utterance, which is amenable to full linguistic analysis. The annotation graphs (as well as transducer approaches [17]), like any graph-based structure, carry no formal explicit means of accommodating the class descriptions (being able to associate an instance of a particular object represented by a graph with the class it belongs to, in order to recover the class information). In particular, the graph representations do not encode the respective production rules (from an appropriate graph grammar) which generated them, while the ETS structs carry all the necessary information for identifying the transformations which led to their emergence.

There are several advantages of the ETS over existing models. First of all, ETS is a scientific formalism rather than an engineering model. The formal basis of the ETS (primitives, transformations, class supertransforms, training and learning algorithms, etc.) is fixed, allowing the researcher to concentrate on a suitable design of the representation. In our case, as outlined in Section 3, this involved the choice of primitives for a small problem dealing with gestural structures of spoken language. Thus, in contrast to the currently popular approaches, this allows for the shift of emphasis to the actual problem at hand (representation and recognition of speech) rather than the recognition and training algorithms, which in speech recognition have become more and more complex over the years. In addition, since the underlying analytic machinery of any ETS representation does not change from one representation to another, this allows for a better and more accurate evaluation of various ETS-based approaches.

Another important advantage, is that the ETS formalism can seamlessly model the emerging perceptual nature of the gestural structures and classes of language, from primitive gestures to the high-level linguistic units: the combinations of gestural structures become more and more syntactically and semantically complex as the levels are ascended in the ETS hierarchy.

5 Conclusion and Future Work

In this paper, we introduced the initial application of the ETS formalism to the low-level linguistic representation of speech based on atomic articulatory gestures represented as ETS primitives. These units encapsulate syntactic and semantic information, which can be reliably derived from real articulatory data using a simple pre-processing algorithm. In the provided example, we showed how the combinations of primitive gestures can be expressed by ETS structs and the recurrent gestural segments of structural history, as the ETS transforms. The groups of structurally related articulations expressed by the gestural transforms are seamlessly encapsulated by the ETS supertransforms providing the means of representing various phonetic classes. We briefly contrasted the ETS approach to speech representation with the currently popular paradigms (both numeric and structural) and described the advantages of the ETS representation over the existing approaches and its promising potential for very flexible structural modelling in speech recognition and linguistics.

The presented work is still in progress and there is quite a lot of room left for both improving the theoretical and practical aspects of the representation. On a theoretical side, we plan to increase the number of classes the system is dealing with, aiming at the experiments on a full-class phone classification problem. This will be followed by the full speech recognition experiments on the MOCHA corpus, which will require the representation to be improved, with the set of the atomic gestures augmented with new physiologically inspired primitives. In addition, we are planning to formalise the multi-level representational approach to speech recognition (the topic only touched in passing in this paper, having concentrated on the initial sensory level only) which will be needed for the full speech recognition experiments. In this approach, the primitive physiological gestures will combine on the higher levels to form the units like syllables and words, which are more complicated than phones. This will allow us to model the hierarchical nature of speech perception, something which can be naturally accomplished within the ETS formalism.

On a practical side, many refinements of the simple pre-processor can be made to improve the gestural detection accuracy. The handling of the EMA trajectories using the algorithm described in [13, 14], is definitely desired along with a better analysis of the EPG data (perhaps following the guidelines of [5]) and the improvements to the pitch detection algorithm on the laryngographic data.

References

1. William Abler, *On the particulate principle of self-diversifying systems*, Journal of Social and Biological Structures **12** (1989), 1–13.
2. Steven Bird and Mark Liberman, *A formal framework for linguistic annotation*, Speech Communication **33** (2001), 23–60.
3. Catherine Browman and Louis Goldstein, *Articulatory gestures as phonological units*, Phonology **6** (1989), 201–251.
4. ———, *Articulatory Phonology: An Overview*, Phonetica **49** (1992), 155–180.
5. Miguel Á. Carreira-Perpiñán and Steve Renals, *Dimensionality reduction of electropalatographic data using latent variable models*, Speech Communication **26** (1998), no. 4, 259–282.
6. Noam Chomsky and Morris Halle, *The Sound Pattern of English*, MIT Press, Cambridge, MA, 1968.
7. Lev Goldfarb, David Gay, Oleg Golubitsky, and Dmitry Korkin, *What is a structural representation ?*, Tech. Report TR04-165, Faculty of Computer Science, University of New Brunswick, Canada, April 2004.
8. Lev Goldfarb, Oleg Golubitsky, and Dmitry Korkin, *What is a structural representation ?*, Tech. Report TR00-137, Faculty of Computer Science, University of New Brunswick, Canada, December 2000.
9. Louis M. Goldstein and Carol Fowler, *Articulatory phonology: a phonology for public language use*, Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities (Antje S. Meyer and Niels O. Schiller, eds.), Mouton de Gruyter, Berlin, 2003.
10. Roman Jakobson, *Six Lectures on Sound and Meaning*, The Harvester Press, Sussex, UK, 1978.
11. Roman Jakobson, Gunnar M. Fant, and Morris Halle, *Preliminaries to Speech Analysis: The distinctive features and their correlates*, MIT Press, Cambridge, MA, 1963.
12. Frederick Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, MA, March 1997.
13. Tzyy-Ping Jung, *An algorithm for deriving an articulatory-phonetic representation*, Ph.D. thesis, Ohio State University, 1993.
14. Tzyy-Ping Jung, Ashok K. Krishnamurthy, Stanley C. Ahalt, Mary E. Beckman, and Sook-Hyang Lee, *Deriving gestural scores from articulatory-movement records using weighted temporal decomposition*, IEEE Transactions on Speech and Audio Processing **4** (1996), no. 1, 2–18.
15. Harold M. Kaplan, *Anatomy and Physiology of Speech*, 2nd ed., McGraw-Hill, 1971.

16. Peter Ladefoged, *A Course in Phonetics*, 4th ed., Harcourt Brace Jovanovich, 2001.
17. M. Mohri, F. Pereira, and M. Riley, *Weighted finite-state transducers in speech recognition*, Computer Speech and Language **16** (2002), no. 1, 69–88.
18. N. Nguyen, *A Matlab toolbox for the analysis of articulatory data in the production of speech*, Behaviour Research Methods, Instruments and Computers **32** (2000), 464–467.
19. Theo Pavlidis, *36 years on the pattern recognition front*, Pattern Recognition Letters **24** (2003), 1–7.
20. Joseph S. Perkell, *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study*, Research Monograph No. 53, MIT Press, 1969.
21. Martin J. Russel and Jeff A. Bilmes, *Introduction to the special issue on new computational paradigms for acoustic modeling in speech recognition*, Computer Speech and Language **17** (2003), no. 2-3, 107–112, (editorial).
22. Michael Studdert-Kennedy and Louis M. Goldstein, *Launching language: The gestural origin of discrete infinity*, Language Evolution (Morten H. Christiansen and Simon Kirby, eds.), Studies in the Evolution of Language, Oxford University Press, New York, 2003.
23. David Talkin, *Robust algorithm for pitch tracking*, Speech Coding and Synthesis (W. B. Kleijn and K. K. Paliwal, eds.), Elsevier Science B.V., 1995.
24. Alan A. Wrench, *A multichannel articulatory database for continuous speech recognition research*, Phonus5, Proceedings of Workshop on Phonetics and Phonology in ASR (University of Saarland), 2000, pp. 1–13.
25. Alan A. Wrench and W. J. Hardcastle, *A multichannel articulatory database and its application for automatic speech recognition*, Proc. 5th Seminar on Speech Production: Models and Data, 2000, pp. 305–308.
26. Steve Young, *Statistical Modelling in Continous Speech Recognition (CSR)*, Proc. Int. Conference on Uncertainty in Artificial Intelligence (Seattle, WA), August 2001.