DETECTION OF SYMBOLIC GESTURAL EVENTS IN ARTICULATORY DATA FOR USE IN STRUCTURAL REPRESENTATIONS OF CONTINUOUS SPEECH

Alexander Gutkin and Simon King

Centre for Speech Technology Research, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW, United Kingdom
{alexander.gutkin,simon.king}@ed.ac.uk

ABSTRACT

One of the crucial issues which often needs to be addressed in structural approaches to speech representation is the choice of fundamental symbolic units of representation. In this paper, a physiologically inspired methodology for defining these symbolic atomic units in terms of primitive articulatory events is proposed. It is shown how the atomic articulatory events (gestures) can be detected directly in the articulatory data. An algorithm for evaluating the reliability of the articulatory events is described and promising results of the experiments conducted on MOCHA articulatory database are presented.

1. INTRODUCTION

Structural representations of speech have received relatively little attention from the speech recognition community. This is due to the fact that purely numerical models are better suited for the task of automatically transcribing continuous speech, reducing the problem to risk minimisation, for which a multitude of efficient numerical algorithms are available [1]. Purely numerical models, however, are not as powerful as structural approaches when the final goal is to discover the structural makeup of the linguistic phenomena (e.g. phones, syllables) under investigation. In the words of Jelinek [2, p. 10]: "These (numerical) 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".

The crucial issues involved in pursuing the structural approach include the choice of an appropriate representational formalism and the choice of features, which for structural approaches play the role of the atomic building blocks. In this paper, we investigate the second issue - the choice of the atomic units of representation. The choice of proposed units, which we call *primitive gestures*, inspired by the physiological combinatorial outlook on speech as advocated by the theory of Articulatory Phonology [3], is discussed in Section 2. We also briefly mention the structural formalisms which have the necessary formal power to seamlessly accommodate such units as the basic building blocks and provide a simple example of gesture-based structural representation. In Section 3, we describe the algorithms which, given the various acoustic and articulatory measurements detect the symbolic primitive gestures directly in the speech data. The procedure for evaluating the reliability of the detection algorithms and corresponding speaker-dependent results for the entire data sets of the two speakers (consisting of 31 minutes of speech data each) from the MOCHA database are presented in Section 4.

A. Gutkin and S. King: Detection of Symbolic Gestural Events in Articulatory Data for Use in Structural Representations of Continuous Speech. In: Proc. IEEE ICASSP, Philadelphia, PA, USA (March 2005),

2. SYMBOLIC ARTICULATORY GESTURES

The main linguistic motivation for this work comes from the theory of articulatory phonology. In articulatory phonology, instead of looking at a shallow description of the act of speech production using traditional units, such as phonemes represented as bundles of phonological distinctive features, 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. 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 [4], thus sharing these combinatorial, *self-diversifying* properties with other natural systems, known from chemistry and developmental biology.

Given the appropriate representational formalism, it becomes possible to talk about the primitive physiological gestures as the atomic information units which combine in various ways to form complex speech patterns. Any satisfactory structural representation of a gesture should encapsulate both the syntactic and the semantic information, thus making the structural combinations of the gestures fully interpretable.

Definition. Informally, a primitive gesture can be defined to be some atomic event (observed in the measurements of the acoustics, articulators and so on) which causes a change in the state of the articulatory organs. Alternatively, it can be seen as the quantum recording of the dynamic interaction between various articulators participating in speech production.

More formally, a primitive gesture π can be represented by a 4-tuple (v, t, E_I, E_O), where v is a vertex of a graph, $t \in \mathbb{R}$ is an internal numeric label, E_I is a set of the incoming edges and E_O is a set of the outgoing edges. Name of the gesture which occurs at time t is provided by v. The elements of the set E_I correspond to the names of the articulators this gesture operates on, plus some additional semantic relations between these articulators, if defined. The set E_O of outgoing edges contains contains the same articulators contained in E_I plus the set of semantic relations and represents the articulators after the occurrence of the gesture. The subsets of E_I and E_O containing semantic relations between the articulators do not necessarily overlap and may be empty.

Such structural entities can be seamlessly represented within the event-based Evolving Transformation System (ETS) formalism [5], where the construction π defined above corresponds to the basic syntactic and semantic unit of the model called *prim-*

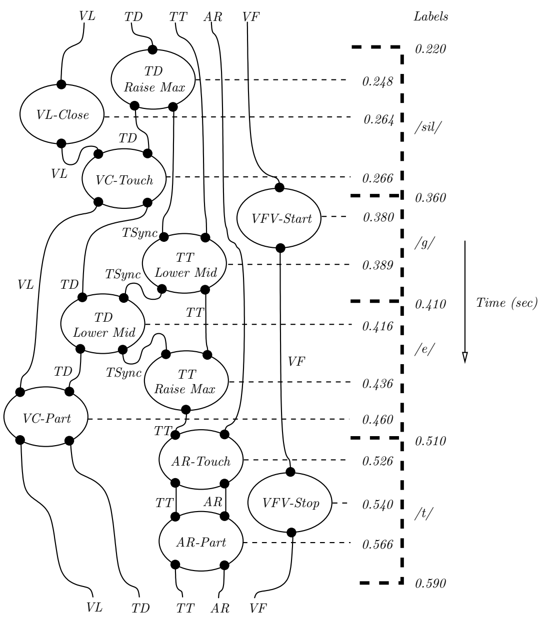

Fig. 1. *Gestural structure of the word "get", constructed using the primitive gestures automatically detected from the data, shown as ETS construction with corresponding phonetic labels.*

itive. The ETS formalism, primarily developed for learning the class representations, encapsulates the procedure for constructing complex structural entities out of the supplied primitives.

Example. Figure 1 shows the gestural structure, expressed within the ETS, for 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 used in this work are given in Table 1. The gestural structure in Figure 1 is constructed on-the-fly from the primitive gestures detected in the available articulatory and acoustic data. For the sake of clarity, only 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. The articulation is achieved by first forming the velar constriction. The constriction is formed by the tongue dorsum TD first rising to its maximum position (TD-RaiseMax) at 0.248 sec, and then completing the constriction before the phone boundary by touching the velum VL (VC-Touch) at 0.266 sec. The constriction is released within the phone 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 the of 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 $/t/$, in the articulation of which the tongue tip (TT), the alveolar ridge (AR) and the vocal folds (VF) participate. The fact that the tongue tip and the dorsum, while being independent, still share some degrees of mechanical freedom, is expressed by the semantic relation (TSync).

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 symbolic atoms standing for primitive gestures.

Table 1. *Articulators involved in the production of primitive gestures and the types of available measurements.*

Organ	Semantics	Measurement Type
UL	upper lip	EMA
LL	lower lip	EMA
UI	upper incisor	EMA
TD	tongue back (dorsum)	EMA, EPG
TT	tongue tip	EMA, EPG
VL	soft palate (velum)	EMA, EPG
HP	hard palate	EPG
AR	alveolar ridge	EPG
VF	vocal folds	laryngeal, acoustic

3. DETECTING THE PRIMITIVE GESTURES

The articulatory corpus used in this work is the MOCHA corpus [6], which 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 (`msak`) and one female (`fsew`) speaker, each consisting of approximately 31 minutes of speech.

Electromagnetic Articulograph (EMA), Laryngograph and Electropalatograph (EPG) measurements were available. The data was automatically labelled using forced alignment. Details on how the measurements were obtained and how the forced alignment was performed can be found in [6].

Given the speech data, in the form described above, and several classes of phonemes, the aim of this work is to detect the articulatory gestures deemed critical (given some theoretical evidence) for the process of articulation of these phonemes. Table 1 shows the articulators we examine together with the types of the corresponding measurements available. Given an articulator (or group of articulators) of interest and the various corresponding streams of measurements, various groups of gestures can be detected, as described below.

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 [7]. 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 [8]. 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 [9], 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 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.

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.

4. EXPERIMENTS

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 `fsew` and `msak` data sets. Since the corpus provides the phonetic labels, it is possible to check whether any of the primitive gestures, *a priori* known to participate in articulations which uniquely define certain phonemes, actually appear during runtime.

Verification Algorithm. 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: (a) The primitive gesture appears within the boundaries (specified by the start and end times) of the phone label currently being processed. (b) The primitive gesture occurs somewhere within the boundaries of several previous phones. In case (b), 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.

Table 2. *Phonemes and constituent gestures under investigation.*

P	N	Primitive Gestures
$/b/$	306	VFV-Start LipsTouch VC-Part AR-Part HP-Part
$/p/$	192	VFV-Stop LipsTouch VC-Part HP-Part
$/g/$	535	VFV-Start VC-Touch TD-RaiseMax AR-Part HP-Part
$/k/$	370	VFV-Stop VC-Touch TD-RaiseMax AR-Part HP-Part
$/d/$	531	VFV-Start AR-Touch TT-RaiseMax
$/t/$	871	VFV-Stop AR-Touch TT-RaiseMax
$/v/$	226	VFV-Start LD-Touch
$/f/$	263	VFV-Stop LD-Touch
$/ng/$	140	VFV-Start VC-Touch TD-RaiseMax VL-Close*
$/m/$	410	VFV-Start LipsTouch VL-Close*
$/n/$	835	VFV-Start AR-Touch TT-RaiseMax VL-Close*
$/ch/$	97	VFV-Stop TT-RaiseMax AR-Touch
$/zh/$	17	VFV-Start TT-RaiseMax HP-Touch
$/sh/$	146	VFV-Stop TT-RaiseMax HP-Touch

Experimental Setup. Table 2 shows the list of 14 consonantal phonemes used in the experiments. For each phoneme, the frequency of occurrence N 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 phoneme. The frequencies of occurrence of the phonetic labels (4939 labels in total) are equal for both male and female speaker data sets. Table 3 provides the description for each of the 13 critical gestures from Table 2. Each gesture is shown alongside the corresponding articulators it operates on, the data stream where the gesture is to be detected and a simple description. For example, the labiodental closure LD-Touch involving the upper incisor and the lower lip is detected in the EMA stream. The name VL-Close* denotes a group consisting of *any* velic aperture gestures resulting in any degree of velum opening, excluding the closure.

The EPG parameters are $\tau_v = 12$, $\tau_p = 6$ and $\tau_a = 9$ for the velar, palatal and alveolar indexes, respectively. These values were determined by manually examining a small subset (two sentences, one for each speaker) of the corpus. The EMA steady state parameter m was set to 10 frames (20 ms for the EMA data sampled at 500 Hz). The number of EMA distance clusters n for all the pairs of articulators in questions was set to 3.

Results. Validation experiments (employing the verification procedure defined above) for each of the 13 critical gestures from Table 3 were conducted on the female and male data sets separately with the results shown in Table 4. The expected frequency of occurrence of each of the critical gestures N_e is the same for the male and the female speaker. For the female speaker, the observed

Table 3. *Primitive gestures critical for the articulation of the phonemes given in Table 2.*

Gesture	Organs	Source	Semantics
LipsTouch	UL,LL	EMA	bilabial closure
VC-Touch	TD,VL	EPG	dorsum touches the velum
VC-Part	TD,VL	EPG	dorsum parts the velum
AR-Touch	TT,AR	EPG	alveolar closure
AR-Part	TT,AR	EPG	alveolar release
HP-Touch	TT,HP	EPG	palatal closure
HP-Part	TT,HP	EPG	palatal release
TD-RaiseMax	TD	EMA	raise dorsum high
TT-RaiseMax	TT	EMA	raise tongue tip high
LD-Touch	TT,UI	EMA	labio-dental closure
VL-Close*	VL	EMA	velum *not* closed
VFV-Start	VF	AC	vocal folds start vibrating
VFV-Stop	VF	AC	vocal folds stop vibrating

Table 4. *Evaluation results for each of the primitive gestures for the female (`fsew`) and male (`msak`) speaker data sets.*

Gesture	N_e	N_o^f	E^f (%)	N_o^m	E^m (%)
LipsTouch	1086	1078	0.74	1079	0.64
VC-Touch	867	803	7.38	750	13.49
VC-Part	676	670	0.89	644	4.73
AR-Touch	2334	2052	12.08	1870	19.88
AR-Part	727	716	1.52	727	0.00
HP-Touch	163	162	0.61	163	0.00
HP-Part	1403	1209	13.86	1325	5.56
TD-RaiseMax	867	854	1.50	844	2.65
TT-RaiseMax	2497	2352	5.81	2388	4.37
LD-Touch	489	479	2.04	481	1.64
VL-Close*	1385	1015	26.71	1086	21.59
VFV-Start	2657	2558	3.73	2573	3.16
VFV-Stop	2282	2213	3.02	2078	8.94
Total	17433	16161	**7.29**	16008	**8.17**

frequency of occurrence of each gesture is specified by N_o^f and the error percentage is given by E^f. For the male speaker, the corresponding measurements are N_o^m and E^m, respectively. As can be seen from the results, while the overall error is reasonably low, some of the primitive gestures are not detected very accurately. The problematic gestures are the alveolar contact (AR-Touch) between the tongue tip and the alveolar ridge (determined from the EPG data), the velar closure (VC-Touch) formed by the tongue dorsum and the velum (determined from EPG data) and the group of gestures (VL-Close*) defining the nasal murmurs (detected in EMA data). The overall error is 7.29% for the female speaker and 8.17% for the male speaker.

5. CONCLUSION AND FUTURE WORK

In this paper, we investigated the potential of the symbolic atoms, which we call atomic primitive gestures, to be used as the basis for the structural representations of the continuous speech, bringing us closer towards the goal of designing a speech recognition system employing articulatory gestures. We have shown how such primitive gestures can be expressed within the ETS, which is one such representational formalism. Within ETS, symbolic gestures, which emerge directly from the articulatory data, combine together in non-trivial ways to form complex speech structures which are fully interpretable and for which efficient learning algorithms are available within the ETS formalism. We have proposed the methodology for detecting the gestures in the continuous speech and shown that the gestures can be extracted with a reasonably low error rate (7.29% and 8.17% error on two speaker data sets of 31 minutes each).

While the results are promising, the presented work is still in progress and there is quite a lot of room for improving the gestural detection accuracy. The modelling of EMA trajectories can be improved by better physiological modelling of the vocal tract making use of the estimated tracing of the hard palate. In order to get more reliable gestural estimates, as well as expand the number of phonetic classes we can handle, the algorithms can be further improved by making more substantial use of the acoustic data, without which it is difficult to infer several classes of sounds (e.g. liquids) based on the articulatory data alone.

6. ACKNOWLEDGMENTS

The authors would like to thank Lev Goldfarb, Mirjam Wester and David Gay for many useful suggestions.

7. REFERENCES

[1] M. J. Russel and J. A. Bilmes, "Introduction to the special issue on new computational paradigms for acoustic modeling in speech recognition," *Computer Speech and Language*, vol. 17, no. 2-3, pp. 107–112, 2003, (editorial).

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

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

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

[5] L. Goldfarb, D. Gay, O. Golubitsky, and D. Korkin, "What is a structural representation ?," Tech. Rep. TR04-165, Faculty of Computer Science, University of New Brunswick, Canada, Apr. 2004.

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

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

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

[9] 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*, vol. 24, no. 7, pp. 881–892, 2002.