Auditory Event Structure and Speech

Fred Cummins

School of Computer Science and Informatics
University College Dublin, Ireland
fred.cummins@ucd.ie

Abstract

Auditory events structure much of the perceived world. Sometimes, two or more sounds are perceived as related, and pertaining to a single event. No well-worked out taxonomy yet exists for auditory events. We consider two-part sounds, perceived as cause and effect (loosely interpreted). Many such sounds occur in ambient environments, and of those, many have suggestively right- or left-headed types of structure. We illustrate these event types, and suggest that familiarity with this type of event structure may motivate their phonologization into familiar structures such as unmarked CV syllable structure. An innate tendency to parse the auditory world into "events" may also facilitate the bootstrapping process of the child language learner. This work is an initial attempt to move the discussion of speech prosody towards a grounding in auditory ecology.

1. Introduction

Are speech sounds *sui generis*? While much effort in phonetics is devoted to describing the structure of speech, and variation in speech patterns across situations and places, rather less effort has been spent in addressing the somewhat perplexing question of why sounds employed in communication have the form they do. Quantal theory [14] represents one well-known attempt to motivate the actual forms employed in speech, and to suggest why these, as opposed to any other forms, occur. In doing so, it starts from the constraint that a symbolic communication system must maintain distinctiveness in the face of often competing articulatory demands, and goes on to demonstrate that some sounds produce well-formed patterns which are relatively insensitive to articulatory variability, thus increasing the likelihood that they be employed in a speech system. Other inquiries into the reason why speech sounds are as they are have focussed on physiological constraints [7, 16] or communicative constraints [8].

Lindblom has emphasized the magnitude of the bootstrapping task facing the infant language learner [11]. Rejecting the notion that children come with the innate ability to extract linguistic features, he argues instead for making minimal assumptions about the initial knowledge of the learner, and motivating phonological structure as an emergent phenomenon. In particular, he suggests that learning may exploit statistical regularities of the speech signal, and sketches an acquisition process that starts with an exemplar-like learning of whole units, which, in later analysis, become divisible and combinable into new forms. We here take up this theme, and suggest that evolution has provided us with auditory systems which are tuned to recognize multiple events as stemming from a single source, and we suggest one possible candidate for a holistic pattern that might be readily recognized by an infant.

We here revisit themes from some older work in the still nascent field of ecological acoustics [6, 5], and consider whether there may exist particular ways of combining sounds such that collectively they are heard as belonging to a single event. In a groundbreaking study from 1984 [15], Warren and Verbrugge identified 'breaking' and 'bouncing' events as ecologically-motivated, well-formed events which had very characteristic structure which was, to a large degree, independent of the detail of the objects involved. Figure 1 shows cartoon versions of spectrograms for bouncing (left) and breaking (right) events. The bounce is characterized by a strong initial impact, followed by a series of impacts with decreasing intensity and temporal separation. The break, on the other hand, has a strong initial event followed by a temporally uncoordinated rush of smaller events, each of which is, itself, a breaking or bouncing event. Again, intensity and interval spacing decrease towards the end of the event.

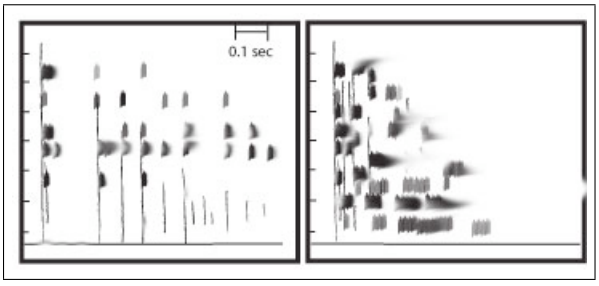

Figure 1: *Cartoon version of spectrograms for bouncing and breaking events (based on [15]).*

While ecological acoustics has continued to extend to address other issues such as the perception of material qualities and information about relative motion, the insight that perceived auditory events may admit of a simplified structural description has not been further developed. In this paper, we suggest an initial characterization of two important event templates which together describe, to an approximation, many events which occur in natural environments. Although each event comprises two distinct sounds, they are heard and understood as emanating from a single event. We provide examples of each and further systematize the characterization of these two event types, such that breaking and bouncing events are both accommodated within our emerging taxonomy. We then pose the question to speech theorists as to whether these event types, so described, may underlie the tendency of languages to favor some forms of complex sound structures, and to disfavor others. In particular, we show that simple CV-structures may be interpreted as unitary events, and speculate that the study of event structure may further our understanding of the role of rhythmic feet in speech.

2. Simple and composite events

Figure 2 illustrates schematic waveforms for four types of single auditory events. Example sounds illustrating each event type are available at [2]. The punctate event might be a gunshot, a slam, a crack, or similar. Examples of sustained continuous sound include an engine running, or the babbling of a brook. Sustained irregular sounds are illustrated by the taps of chalk on a board or the crack of a gun battle, while sustained periodic sounds include footsteps and hammering.

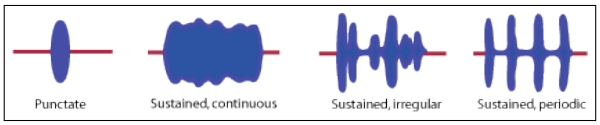

Figure 2: *Four simplified continuous sound event types.*

More interesting than single sound types are composite events. At this stage, we do not consider events with more than two components. Consider the two event types described in [15]. Each can be schematized as a punctate event (the impact) followed by a modulated continuous event: periodic for the bounce and irregular for the break. The modulation required includes a progressive decrease in amplitude and in inter-event intervals. Caricatures are shown in Fig 3. This simplification suggests a template-like approach to the characterization of composite sound events. For example, the sound of a match striking can be described as a punctate event followed by a sustained, suitably modulated, continuous sound. A branch breaking has a punctate crack, followed by a tearing of decreasing intensity.

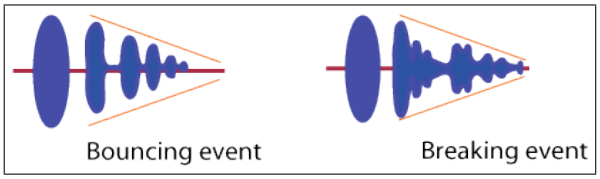

Figure 3: *Breaking and bouncing events reconsidered as simple event structures.*

We have now identified one kind of two-part event, with a strong head (the punctate event) and a following sustained sound of decreasing intensity (and possibly with decreasing temporal and frequency characteristics). Current work in our laboratory is investigating whether these events are perceived as wholes because they are perceived as stemming from cause-and-effect relations, much like certain properties of simple visual stimuli have been shown to induce an obligatory impression of causality [12, 13].

Interestingly, we can reverse the structure we have just identified, by placing a suitably modulated continuous sound before a punctate event. In this fashion, we can arrive at a reasonable structural description of a drawer or screen door closing, a skid and crash, and numerous other simple events. Indeed, closing and opening events may very often display complementary event structure, and may constitute useful areas of inquiry in the manner of [15]. Two real events are illustrated in Figure 4, where it can be seen that these physically very different event types share structural properties, which are here mirrored along the time line.

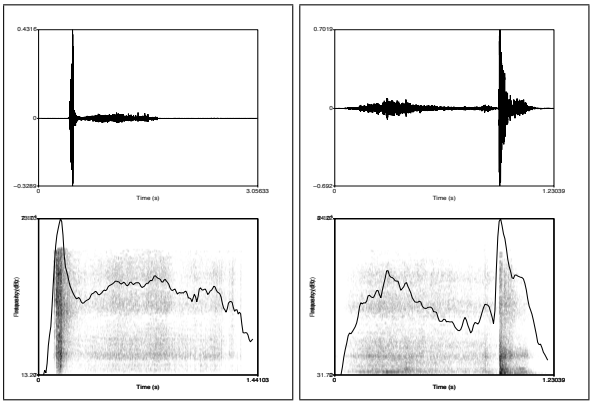

Figure 4: *Waveforms, spectrograms and intensity contours of a match being lit (left) and a drawer closing (right).*

We are currently synthesizing artificial sound events with structures based on these two forms [2]. Our ongoing work asks whether the perception of causality may be stronger when the punctate event is the first element, producing a stronger impression of causal linkage between the two sounds. There is clearly much work to be done here in uncovering the relationship between complex event structure and the perceptions thereby induced.

3. The use of sound in language

Although this work is at an early stage, the goal of systematizing the description of auditory events raises some interesting questions for phoneticians and phonologists.

It is well known that language may be supported by speech, signing, writing, and other media. In this respect, the use of sound is not central to language, but should be seen as a medium within which a system of sufficiently systematic contrasts can be built. The perceptual system has evolved to be of use in real world, complex, environments, in which certain sound events are of potential importance to an organism. In particular, if a single event gives rise to multiple sounds, it is clearly of benefit to the organism to perceive their common source, rather than a set of disconnected sounds. This is not unlike the related case in vision, where objects are perceived as whole entities, and not as collections of visual properties such as color, texture, etc.

The hypothesis being offered for consideration here is that certain structural features of complex sounds may strongly affect their perception as a single event. This is a rather abstract notion of an ”event”. The specific example of a breaking event is characterized by overall temporal and intensity dynamics, and not by the spectral detail that would distinguish a crystal vase from a porcelain jug. If we can identify canonical dynamics which specify event types, it may be possible to relate these structures to more complex and adaptive uses of sound, e.g. in speech.

In an influential paper, Fowler [4] proposed a direct-realist theory of speech events. In a programmatic agenda in the spirit of the ecological psychology tradition within visual science, she proposed that speakers produce speech events which are ”phonetically structured articulations”, and that these articulations, in turn, are perceived by listeners. The goal of the direct realist approach was to account for direct immediate perception of real world events of relevance to an organism, without re-

course to putative disembodied cognitive processes of inference or hypothesis testing. The articulations understood to be perceived were more or less those which constitute the atomic units of Articulatory Phonology [1]. One problem which immediately raised itself is that untrained listeners are not aware of the perception of articulation, and are unable to report constituent gestures of an utterance. This contrasts with vision, where no schooling is required to perceive a table or a duck, and the perception is clearly of the distal object.

Despite the appeal of a direct realist approach to speech perception, this apparent failure of listeners to easily recover articulations has led to little further work in this direction[1]. The hypothesis presented here suggests that there might, instead, be a rather more abstract notion of 'event' which is of ecological significance to listeners and which may better account for the facility with which the speech stream is readily parsed, at least in the initial stage in which linguistic features are being discovered. As will be illustrated below, there are several frequently occurring sound patterns in speech which seem, at first blush, to bear similarity to the simple two-part events described above.

4. Phonologization

Speech appeared relatively suddenly upon the world's stage. Although very little is known about the genesis of language, we do know that it appeared within a time span that precludes the evolution by conventional means of any substantial biological structures or processes. Speech is not built of nothing; rather, the constituent elements and the means with which they are sequenced, transmitted and perceived, are necessarily built upon the basis of motor control and perceptual apparatuses which evolved over a much larger period, and which primarily serve other purposes.

These observations suggest that if the auditory system had developed the ability to identify and directly perceive individual distal events which give rise to sequential, disparate sounds, this *modus operandi* would be potentially available for use in the production and perception of speech. In this characterization, it is not "articulatory events" per se which are perceived, but rather "events", characterized by macroscopic temporal dynamics of a kind with other, non-speech, events. The events within the speech stream are neither *sui generis* nor entirely like other sounds from the environment, but they maintain sufficient structural coherence to be readily parsed by a perceptual system which specializes in seeing through component sounds to their unitary underlying events.

Of course languages are not built of unformed clay either. If there is a pre-existing ability to parse the acoustic environment into events based on macroscopic dynamics, then this ability becomes a possible source of exaptation within a linguistic contrastive system. Events would thus provide raw material for initial construction of a parseable stream of sound produced almost invisibly within the vocal tract. The exact nature of the underlying event (the articulatory gesture or gestures) is not the important element here, and need not be faithfully perceived as an articulatory event. What is important, is that the sound so produced exhibit some of the basic dynamic structure which specifies events in the world, thereby facilitating the parsing of the acoustic stream.

[1] A close relative of the direct-realist approach is the Motor Theory of Liberman and Mattingly [10]. This theory has had rather more exposure of late, in part due to the discovery of mirror neurons which suggest a direct neurophysiological counterpart to the basic elements of the theory

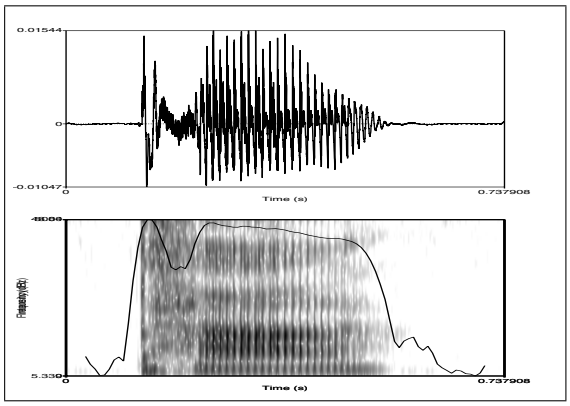

Figure 5: *Waveform, spectrogram and intensity contour for a single utterance of /pa/.*

Figure 5 shows the structure of a single utterance of the syllable /pa/. The macroscopic structure of this utterance bears strong resemblance to Fig 4 (left) and to the template structures in Figure 3. The stop release is a distinct punctate sound followed by the sustained voicing of the vowel. CV is of course a maximally unmarked syllable type, and a plosive consonant in the C position is also unmarked, lending some credence to the suggestion that this basic structure may reflect a form of auditory event which perceptual systems are already attuned to before integration into a speech system. Modelling work has suggested that a CV structure is more readily perceptible than VC [9], but the reasons for this are unclear. Infants at about 7 months begin producing canonical syllable structures, frequently reduplicated, suggesting that the searching behavior of babbling is geared towards the discovery of coordination patterns which give rise to event-like sounds. In the present account, the CV structure appears likely to be an unmarked unit in the speech stream as it exhibits a dynamic structure common to very many ecological events, and thus represents a readily parseable unit in speech.

The commonality between speech structure and event structure beyond speech may be of importance both in accounts of how speech systems came into being in the first place, and how infants begin to parse the speech stream into individual components. In each case, what is suggested here is that the innate predisposition to attribute suitably formed sound sequences to a common event source may be exploited by the speech system in establishing a sound system capable of supporting a sufficiently rich set of categorical distinctions. In the development of speech systems, the bias towards fusing conjoined sounds into singular perceived events provides a starting point for the development of a combinatorial sound system. From an ontogenetic point of view, this disposition may allow an initial parsing of the speech stream, thereby providing a handle on the daunting task of uncovering phonological structure from scratch.

5. Discussion

We have speculated that auditory event structure may play a determining role in providing the raw material out of which speech is built. Much remains to be done in discovering which properties of composite sound sequences cause them to be perceived as stemming from a single event.

The CV syllable appears as one obvious example of a recur-

ring structure which is privileged in speech and which seems to be potentially grounded in a phylogenetically old ability of perceptual systems to parse specific kinds of sound sequences into constituent events, rather than individual sounds. Until further work is done in identifying those characteristics of groups of sounds which favour their interpretation by a perceiver as stemming from a single event, it will not be possible to extend this initial observation much further. However, a systematic investigation of auditory event sound structure may provide novel means for understanding the predominance of some forms at the expense of others across the world's languages. It may also suggest how it is that infant learners manage to parse the speech stream *before* they have acquired any knowledge of phonological contrast or linguistic feature.

In the absence (for now) of such work, we might nevertheless venture to look ahead, and suggest further lines of inquiry. The Principles and Parameters approach to metrical structure in language was well outlined in [3], where the systematic analysis of stress systems across many languages suggests that language learners need only identify the setting of a small number of parameters in order to learn the stress system of any given natural language. These principles include such examples as "Feet are built from the [Left/Right]" and "Feet are strong on the [Left/Right]". This approach is admirable in its parsimony, suggesting that, given the right parameters, the learning task in this instance is tractable. But where do such parameters come from? What is the innate knowledge expressed in this theory as "Principles"? On the current account, both right-headed and left-headed combinations of sounds specify common event structures (though there may turn out to be an innate bias towards left-headed events). The left-headed structure captures the combination of sounds in a CV-sylable well. Perhaps a sensitivity to these structures may pre-dispose a learner to pick out rhythmical units in a language which employs stress feet.

The approach being proposed here suggests that speech sound patterns are extrapolations of sound structures found in the acoustic environment and specific to particular kinds of events. The structures we have sketched out above include both proto-iambic and proto-trochaic patterns. If these template event structures can be shown to be effective in inducing the perception of a single event in listeners, then it suggests that the knowledge that some salient units of speech can be right or left headed, where the head is a relatively strong element, might indeed be grounded in a phylogenetically older form of knowledge about the sound world which surrounds us. Speech may be special, but perhaps we phoneticians sometimes overlook commonalities between speech and other ambient sound information.

6. References

[1] C. Browman and L. Goldstein. Articulatory gestures as phonological units. *Phonology*, 6:201–251, 1990.

[2] F. Cummins. Auditory event structure. http://cspeech.ucd.ie/∼fred/research/auditoryeventstructure.

[3] B. E. Dresher and J. D. Kaye. A computational learning model for metrical phonology. *Cognition*, 34:137–95, 1990.

[4] C. A. Fowler, M. R. Smith, and L. G. Tassinary. Perception of syllable timing by prebabbling infants. *Journal of the Acoustical Society of America*, 79(3):814–825, 1986.

[5] W. W. Gaver. How do we hear in the world?: Explorations in ecological acoustics. *Ecological Psychology*, 5(4):285–313, 1993.

[6] W. W. Gaver. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology*, 5(1):1–29, 1993.

[7] K. S. Harris. Vowel duration change and its underlying physiological mechanisms. *Language and Speech*, 21(4):123–130, 1978.

[8] J. M. Iverson and E. Thelen. Hand, mouth and brain: The dynamic emergence of speech and gesture. *Journal of Consciousness Studies*, 6(11–12):19–40, 1999.

[9] M. F. Joanisse. Exploring syllable structure in connectionist networks. In *Proc ICPhS 99*, pages 731–734, 1999.

[10] A. M. Liberman and I. G. Mattingly. The motor theory of speech perception revised. *Cognition*, 21:1–36, 1985.

[11] B. Lindblom. Developmental origins of adult phonology: The interplay between phonetic emergents and the evolutionary adaptations of sound patterns. *Phonetica*, 57, 2000.

[12] A. Michotte. *La perception de la causalité*. Publications Universitaires de Louvain, 1954.

[13] B. J. Scholl and P. D. Tremoulet. Perceptual causality and animacy. *Trends in Cognitive Sciences*, 4:299–309, Aug. 2000.

[14] K. N. Stevens. On the quantal nature of speech. *Journal of Phonetics*, 17:3–45, 1989.

[15] W. H. Warren and R. R. Verbrugge. Auditory perception of breaking and bouncing events: A case study in ecological acoustics. *Journal of Experimental Psychology: Human Perception and Performance*, 10:704–712, 1984.

[16] Y. Xu. Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, 33:319–337, 2001.