An Embodiment Paradigm for Speech Recognition Systems

Gina Joue and Julie Carson-Berndsen
Department of Computer Science
University College Dublin
{ gina.joue, julie.berndsen }@ucd.ie

Abstract

The problems of conventional speech recognition approaches include incomplete linguistic knowledge and inability to deal with underspecification. These issues can be addressed by understanding the constraints of speech to predict speech tendencies. We believe that understanding what constraints exist requires an embodied view of speech and that the traditional disembodied view of speech is the fundamental limitation on the robustness of many speech systems. We argue that viewing speech as a form of embodied cognition, or within context of its production and use, provides important insights in speech structure and speech recognition. In making this claim, this paper briefly outlines a strongly embodied account of cognition and develops from that an embodiment paradigm for speech recognition. The embodiment paradigm proposed leads to both an explanatory and descriptive account of linguistic structure. It simplifies the view of speech structure for automatic speech recognisers, by considering only the most directly relevant motivations or constraints influencing communication and thus speech.

1. Introduction

Conventional speech recognition approaches have predominantly treated the speech stream as simply a physics problem and ignored the importance of many of the different factors influencing speech structure. These systems have also generally bypassed phonological modelling and relied on pattern matching algorithms to process the speech signal and oversimplistic phonological rules and/or pronunciation dictionaries to deal with pronunciation variants. The conventional speech technology techniques may be powerful and efficient, but they have proved limited in coping with the variabilities in natural speech. We argue that this disembodied view of the problem of speech recognition is what limits these systems, and thus present an embodiment paradigm for speech.

The general embodied cognition approach claims that cognition can be better understood as a continuous interaction of the nervous system, the body, and the environment. The body is the necessary interface between internal processes (such as mental processes) and the environment. Some traditional cognitive scientists have supported the more embodied study of cognition, but the main school of thought continued to view human cognition only as purely symbolic processes and internal representations in the head (for example), and dissociated from any interaction with the environment. The embodied view only recently gained more influence (cf. [1]) in the realisation that traditional approaches ignored important aspects of cognition.

The term embodiment, however, has come to loosely describe anything that acknowledges the use of the physical body (such as: the eye is used in vision), which we will refer to as weak embodiment. In this paper, we advocate a strongly embodied account for speech. Note that weak and strong embodiment have appeared as polaris of different scales in the cognitive science literature. In this paper, strong embodiment based on the more general overview that the physical (body) experience (such as the involvement of articulators speech production) is not only important in cognition, but that the body and its interaction with the environment are necessary in understanding cognition. Specifically for speech, a weak embodiment of speech would claim that the physical experience and its interaction with the context of speaking, such as social context, are necessary to understanding speech structure and its use.

Weak embodiment can be seen in some existing theories of speech perception, phonology and methods in automatic speech recognition systems. Motor-based perception theories, such as the one put forth by Liberman and Mattingly [2], highlight the interdependency of physical motor actions and the nervous system. Lindblom’s Hyper-Hypo theory [3] and Nearey’s Double-weak theory [4] of speech perception emphasise how both perceptual and articulatory/muscular constraints shape speech.

In phonology, the importance of strong embodiment is only weakly addressed. Traditional generative phonology only viewed language as internal symbolic manipulation; thus, ignoring embodiment altogether. Articulatory phonology [5] only admits to a weak embodiment as it internalises the physical (articulatory) gestures and closely relates execution (physical) and planning (internal) of speech. Perhaps closest to the strong embodiment of speech in phonology, is Phonology as Human Behavior [6, 7], developed in the 1960s at Columbia University.

Although automatic speech recognition systems borrowed general ideas from linguistics, most mainstream recognisers do not exploit the insights afforded by phonological theories. Only a few non-mainstream systems have taken such an approach. The speech recognition systems developed by Deng and colleagues [8] borrowed from linguistic theories such as articulatory phonology and seem to support a weakly embodied paradigm for speech technology and even make assertions supporting a strongly embodied paradigm. Boersma’s speech system [9], influenced by functional phonology and partially by Phonology as Human Behavior, also straddles the embodied speech paradigm. However both these approaches utilise more knowledge or parameters than seems necessary because their embodied accounts were not strong enough.

The next section details more closely at what defines a strongly embodied account of speech. The realisation of this embodiment paradigm in Phonology as Human Behavior is
discussed in the Section 2, and the paradigm is further applied to speech technology, in a computational framework LIPS in Section 3.

2. The Embodiment Paradigm for Speech

Knowledge can be computationally expensive for speech technology applications, which is partly why data-driven statistical techniques are so efficient. Understanding the constraints involved in speech and its structure compensates for having incomplete or partial knowledge. Understanding what constraints exist requires an embodied view of speech.

Speech as an embodied form of cognition takes into account how speech is used in a social context (the environment), the physical parameters or constraints involved in speech (articulatory and auditory factors), and the interaction of the environment and physical constraints. This can be alternatively stated in terms of general human problem-solving techniques, where there are goals, constraints, and the tendency to find efficient methods to reach the goals despite the constraints. As these constraints directly influence speech, to ignore them undermines the understanding of the structure of language and the means to explain the vastly predictable commonalities across human languages (and even the differences between languages).

In not treating language as an embodied form of cognition, nothing then constrains the differences among languages and what/how languages do develop. Linguistics and the success of data-driven recognition systems do tell us that there are consistent and systematic similarities among languages. Acoustic phonetic studies such as Kawasaki-Fukumori’s studies on phonotactics [10] also imply that universals in speech structure can be explained by an embodied account of speech. All languages’ sound systems and phonotactics are influenced by the same human constraint factors that arise in attaining the goal of maximal communication. Differences among languages merely are differences in how individual languages weigh the local and global constraints of speech, i.e. how the interactions of these constraints manifest in the language.

The weakly embodied theories and approaches mentioned in the previous section can be seen as only dealing with local goals and their respective local constraints. The fully-embodied, or strongly embodied, account that we propose takes into the focal point both levels (local and global) of goal-orientation. Unlike weak embodiment accounts, a strong embodiment account by default reduces the amount of parameters needed to explain speech structure by considering the comprehensive whole rather than immersing in the infinitely possible number of details that can be culled from the acoustic speech stream.

There are two types of goals in speech: communication (environmental) and articulatory (physical body) intentions (whether they are targets, schemas, or windows). These types of goals can be seen as hierarchical local and global goals, where articulatory targets and auditory capability are the local goals which need to be achieved in order to fulfill the global goal of communication. That is, attaining the local goals are constraints on attaining the global goal. We can thus see constraints on both a local and global level: the global constraint is the interaction among sets of local constraints of local goals. Local constraints are the interactions within a set of constraints for each local goal.

As mentioned, the global goal for speech is to maintain maximum communication. The local goals are:
- easy learnability (infants have to be able to learn the language effectively)
- easy decoding (understanding without constant strain on the human physical system, as is dependent on the nature of human physiology, in both optimal decoding conditions and in less optimal or noisy conditions)
- easy encoding (without constant physical strain)

To summarise in terms of constraints, speech as an embodied cognition entails the following:
- the GLOBAL constraint is the interaction of all local constraints involved in achieving the goal of maximum communication
- the LOCAL constraints include:
  - human physiological constraints
    - execution constraints
    - perception constraints
    - learnability constraints
  - environmental constraints (social context, noise context)

Figure 1 is an entity-relation diagram illustrating these factors that influence speech.

The achievement of the global goal of maximum communication depends on the interaction among the local constraints of the goals of learnability and the ease of encoding and decoding of the tool used for communication. Speech, as a tool for communication, inherits the dependencies of the goal of communication. In context, this environment for speech is the human social environment where human physiology constrains decoding (auditory, visual, memory limitations) and encoding (articulatory-muscular, memory limitations). Although the entities are separated out in Figure 1, they are interrelated and interdependent (as is indicated by the one-to-one connections among them). For example, articulatory ease is partly dictated by learnability as well as familiarity (both subsumed in easy decoding). However, aspects that involve high learnability never reach familiarity.

![Figure 1: Entity-Relation Diagram of Factors Influencing Speech Structure](image-url)
The view of strong embodiment, which emphasises goal-oriented behaviour and constraint interaction, provides both an explanatory and descriptive account of speech sound structure (in other words, phonology). It simplifies the view of speech structure (phonology) for rather than using detailed neuromuscular and articulatory parameters, used in weak embodied approaches and which are more complex than necessary, we need only to consider the higher-level parameters (motivations) that are most relevant to speech in influencing communication.

3. An Embodied Speech System

The last section pointed out characteristics of the knowledge and methods needed for an embodied approach to speech. This section discusses the representations and techniques which have those characteristics, allowing us to build such a system.

Phonology as Human Behavior [6,7] posits that the apparently motivated distribution of speech sounds results from the interaction of factors illustrated in Figure 1. The phonology has constraints both on the local and global goal levels. It views the interaction of local goals (what Tobin terms the Human Factor) and the global goal (termed the Communication Factor) as a synergistic relationship. In Phonology as Human Behavior, the local level constraints are described on the basis of phonological features. For example, features include the active articulators used to produce speech sounds, which correspond to execution constraints, and perceptual salience of given speech sounds, which are auditory constraints in speech. The global level constraints are (dis)favourings which describe speech structure. Examples of global level constraints would include the interaction of articulatory-muscular coordination and/or control difficulties with perceptual and learnability ease of a given sequence of speech sounds.

We have previously proposed that incorporating the factors highlighted by Phonology as Human Behavior into an automatic speech recogniser provides the system with a predictive and diagnostic advantage (cf. [11]). That is, the system would be able to handle underspecified structures as is common in natural speech, as well as evaluate its hypotheses and generate most likely alternative ones. We emphasise that the compatibility of the strong embodiment of speech paradigm with our approaches is the strength of our system.

The implication of the strong embodied view of speech is an explained predictability of speech structure. Constraints derived from this view allow for the possibility of coping with processing speech in speech recognition systems without complete phonological knowledge or thorough detailed understanding of human speech processing knowledge. It allows for flexible judgements needed to cope with variability in real-life speech.

As we are currently only dealing with the acoustic stream, we need to be able to process it in a manner which preserves information important to the factors influencing speech structure. In our system we utilise the Time Map model [12, 13] as realised in the Language Independent Phonotactic System [14].

LIPS has been designed as a development environment specifically to address issues of robustness in speech recognition. The Time Map model uses a complete description of the phonotactics of a language together with axioms of event logic to interpret multilinear representations of speech utterances. The phonotactics are described in terms of a finite state automaton, the transitions of which define constraints on overlap relations between features in the multilinear representation. If the input representation is underspecified, the constraints on the transitions may be relaxed according to a constraint ranking which is based on a functional cognitive analysis with respect to the features of Phonology as Human Behavior. Using the phonotactic constraints together with their rankings, the output representation may also be extrapolated, augmenting the representation by features which were missing in the input.

LIPS preserves and highlights the parameters for an embodied approach to speech recognition, providing a compact computational framework for a strongly embodied speech recognition architecture. This architecture is depicted in Figure 2. Unlike conventional speech models which treat speech as sequential concatenated units, Time Map phonology is based on a more psycholinguistically-sound approach. It processes speech as multilinear tiers of phonological events, which is important to preserve descriptability of local constraints in speech. This approach models speech affects such as coarticulation, which represents a local constraint on articulation, as temporal relations of speech events (features or parameters). As the LIPS system uses phonological constraints, in terms of feature-based phonotactics, to incorporate native speaker’s linguistic intuitions, it eases coping with new words and noisy speech. On this level, it incorporates a hybrid interaction of local constraints with some global constraint, since a speaker’s linguistic intuition is based on probability judgements of global level constraints. That is, it incorporates the likelihood of a sequence of sounds to be in a language, given the speaker’s knowledge of the language (how local constraints have played out) and the communicative load of the sounds (global constraints).

![Figure 2: Architecture for syllable recognition based on the embodiment of speech paradigm](image-url)
4. Conclusions

This paper detailed a strong embodied account of speech. It argued that an embodied view of speech provides insight into phonological analyses and enhances the robustness of automatic speech recognisers by incorporating the ability to make predictions of tendencies in speech sound structure. A speech recognition architecture based on the embodiment of speech paradigm was also presented. Although multimodal speech recognition is a further step to an embodied paradigm for speech recognition, we wanted to highlight that the strong embodiment paradigm does not merely superficially instantiate the use of the body (in case of speech, hand gestures, facial expression, and oral articulators) with the less obvious example of audio-only speech recognition.

5. References