Speech Recognition with Syllables and Concepts

Paul De Palma
Department of Computer Science
Gonzaga University
Spokane, WA, USA
depalma@gonzaga.edu

Charles Wooters
International Computer Science Institute
Berkeley, CA, USA
wooters@icsi.berkeley.edu

***Abstract*--The transformation of speech to words is not necessary to emulate human linguistic performance in some contexts. A large vocabulary continuous speech recognition (LVCSR) system can perform more accurately with a syllable-based, as opposed to a word-based, language model. This accuracy can be further enhanced by the addition of a concept model, where a concept is an equivalence class of words and phrases.**

Index Terms—**speech recognition, syllable, concept**

I. INTRODUCTION

We began our work with the observation that spontaneous speech is not simply unedited writing. The differences between speech and writing have been treated at length in the literary critical/linguistic/psychological literature. They include a wealth of counter-intuitive examples. For instance, utterances like "John bought a motorcycle," in which full nouns designate all participants in an utterance are rare in spontaneous speech [16, p.4]. As Miller and Weinert point out at the very beginning of their book, "syntactic structure of phrases and clauses in spontaneous spoken language is very different from the structure of phrases and clauses in written language," and, "the organization of spontaneous spoken discourse is very different from the organization of written discourse" [9, p. 1].

Further, these structures are cross-linguistic. One of the more interesting aspects of this distinction for language study is that the generative tradition--usually identified with work of Noam Chomsky, along with anthropological linguistics, relies heavily on elicited data. That is, the investigator provides his or her own examples of a structure under consideration or elicits a structure from an informant, as illustrated in this passage on field methods from *The Linguistics Encyclopedia*: "The fieldworker … already has the word for *eye* so s/he asks for *two eyes*" [4, p. 163]. This is contrived in the latter case, as the author suggests, and in the former case simply odd. There the linguist is playing the role of both observer and observee. In both cases, however, the distinction between real, living language, language as it is used (and spoken), has been erased.

Though we assume that the many researchers who developed automatic speech recognition understand quite well the uncontroversial observation that speech and language differ substantially, that understanding is sometimes not obvious in the work itself. Here is how the National Institute of Standards and Technology describes one of its projects: "The Rich Transcription evaluation series promotes and gauges advances in the state-of-the-art in several automatic speech recognition technologies. The goal of the evaluation series is to create recognition technologies that will produce transcriptions that are more readable by humans and more useful for machines" [11]. Even if we allow that the project under discussion is transcription itself, all one needs to do is to examine any "correct" transcription against which an automatic speech recognition system is scored to find a print bias. They sometimes—not always, of course—read like the interviews found in newspapers and magazines. The following passage is taken from the New York Times: "We have never seen anything like this in our history. Even the British colonial rule, they stopped chasing people around when they ran into a monastery" [13]. Who could argue that the reporter has transformed an acoustic signal into words? Though we might wish for a recording of the interview, we can get a sense of how it must have sounded by looking at the very first segment of the Buckeye Corpus.

> yes <VOCNOISE> i uh <SIL> um <SIL> uh <VOCNOISE> lordy <VOCNOISE> um <VOCNOISE> grew up on the westside i went to <EXCLUDE-name> my husband went to <EXCLUDE-name> um <SIL> proximity wise is probably within a mile of each other we were kind of high school sweethearts and <VOCNOISE> the whole bit <SIL> um <VOCNOISE> his dad still lives in grove city my mom lives still <SIL> at our old family house there on the westside <VOCNOISE> and

we moved <SIL> um <SIL> also on the westside
probably couple miles from my mom.

Now, the obvious rejoinder is that the snippet from Buckeye, through a transcription of speech, is, in fact, text. If we could get a speech recognizer to reproduce transcriptions of speech accurately, then we could count the decades long project of automatic speech recognition a success. Perhaps so. But we would like to do more. In fact, we derive the inspiration for our research from the way that humans process speech, a claim that we will return to. For now, we address what must be the next objection. A speech recognizer is an engineering artifact. It need not duplicate the functional characteristics of the system it emulates—human beings, in this case—just as air planes need not flap their wings. But there is a counter-objection. While, aeronautical engineers and birds have different goals. An LVCSR system and a court reporter, for example, have remarkably similar goals [5]. Perhaps we can learn something from the only system that has successfully solved the speech recognition problem, humans themselves.

For starters, it appears that humans are unable to produce a text transcription of a conversation. Though we do quite a nice job of recognizing speakers from just a few seconds of telephone speech [14], producing a text transcription--not a summary, not a sanitized version, but an accurate text transcription—seems a more difficult, possibly inhuman task. The literature overflows with studies of short and long-term memory, but very few of the studies appear to ask a random sample of subjects to recall genuine human speech. Nevertheless, the literature indicates that verbal memory capacity appears to bear an interesting relationship to speech production. We seem to be able to retain as much acoustic data as we can pronounce in 2 seconds [1]. While we would never suggest that AI research limit itself to human performance, we hypothesize that there is much to be gained by building a recognizer that could output something other than a word string. To see this, one has to imagine a recognizer that handles spoken plane reservations. Passing the sense of the caller's speech to a domain knowledge system would be the first step in solving a problem of significant scope.

The next step is to clarify what we mean by "something other than a word string" and "sense of the caller's speech?" These reduce to syllables and concepts. At first blush, syllables are principled subdivisions of a word, and a concept is an equivalence class of words and word strings that seem to mean more or less the same thing [2], [3]. Conventional automatic speech recognition (ASR) generates a word string given a sequence of acoustic observations. Instead, we begin by generating a syllable string. Notice that we have reduced the search space: in English—though the number is far from certain—there are fewer syllables than words. Of course, this syllable string is not useful by itself. We propose instead to probabilistically map it to concept strings. We hypothesize that there are fewer concepts than words, since we define them as words and collections of words. We call this hypothesized dual reduction in search space the Symbol-Concept Hypothesis (SCH). SCH claims that the reduced search space will result in more robust recognizers. Though SCH can be argued using the axioms of probability, in the final analysis, it is an empirical hypothesis and must be demonstrated experimentally. This brings us to our research, a four-phase, multi-year effort:

- Phase I: Gather preliminary data about SCH using ATIS0 and another small corpus
- Phase II: Reproduce the syllable results from Phase I using both the TIMIT and ATIS2 corpora
- Phase III: Build a probabilistic concept generator and a probabilistic concept model
- Phase IV: Pass results from Phase III to a domain knowledge system and from there to a speech synthesizer that generates a response.

II. THE UNDERAPPRECIATED SYLLABLE

The complexity of the syllable appears to be underappreciated in the literature on speech recognition. Rabiner and Young are not atypical when they provide only two index entries for *syllable*, and appear to regard it as an easily identified sub-word unit [12]. As we have previously noted, this is peculiar since the number of purported syllables in English varies by a factor of 30, depending upon whom one reads [3]. The certainty surrounding discussion of a syllable would surprise a linguist. Kohler, for example, drew attention when he denied its existence as a linguistic universal in the mid-sixties [6]. And, as the widely varying number of English syllables reported in the ASR literature suggests, the syllable is especially tough to pin down as a phonological element, partly "because it is not a sound, but rather an abstraction over particular organizations of sound" [2, p. 43]. Nevertheless, most linguists, especially those in the generative tradition claim that syllables are important units of phonological structure.

Much of their importance appears to be their usefulness in deriving compact phonological rules. Consider the sound sequences [t r] and [t l], the first is found word initially in English while the last is not. This is a rule. But the introduction of the syllable lets us be more precise. Consider what happens with these phone sequences when they occur word medially, as in *Atlantic* and *atrocious*? The /t/ in *atrocious* is aspirated, but the /t/ in *Atlantic* is glottalized. If argue [t r] can be a syllable onset while [t l] cannot, we can put syllable breaks before /t/ in *atrocious* and after /t/ in *Atlantic.* Now we can gather the aspirated and glottalized /t/ allophones into a generalization: syllable-initial /t/ is aspirated and syllable-final /t/ is glottalized. The notion of a syllable lets us express phonotactic generalizations that might otherwise go unnoticed.

The structure of a syllable as a unit also appears to exhibit regularity. Linguists have argued for a hierarchical structure since the forties. In this view, the syllable is composed of an optional consonant onset followed by a rhyme, which is further divided into a nucleus followed by an optional consonant coda. Figure 1 gives the basic form. Speech sounds are often ranked by amplitude, with the following phone classes ranging from highest sonority to lowest: vowels, glides, liquids, nasals, and obstruents. For example, [a] is more sonorous than [w], which is more sonorous than [l], which is more sonorous than [n], which is more sonorous than [t]. The basic syllable has a low sonority consonant onset followed by a high sonority vowel in the nucleus position. {V, CV} seems to be a part of the syllabic inventory in all languages. Many add codas to these basic forms, producing {V, CV, VC, CVC}.

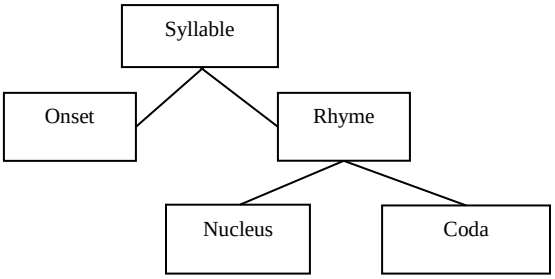

Fig. 1. General Syllable Structure

Languages with complex syllable inventories, like English, supplement the basic inventories in ways that are mostly consistent with the sonority hierarchy. So, in English we have double consonant onsets (*twist*, *priest*, *fly*, *gripe)*. If we think of these words as being composed of a nucleus at the vowel, in every case, the double consonant onset, has a consonant lower in the sonority hierarchy preceding one above it in the hierarchy. So, in *twist* obstruent [t] precedes glide [w]. Sonority rises at the vowel [i], falls at the fricative [s], an obstruent, and falls further at the final obstruent [t], which is, by the way, lower in the sonority hierarchy than [s]. The sonority hierarchy in English is useful but not absolute. For instance, the obstruent [p] does not combine with the glide [w], though the sonority hierarchy predicts that it would. English seems to avoid grouping sounds that use the same articulators, the lips in this example. Even this is not absolute. The fricative [s] is serial offender. It violates the constraint on adjacent consonants using the same point of articulation (e.g., *slip*), violates the sonority hierarchy (e.g., *spit*), and even initiates triple consonant onsets (e.g., *split*). And what about *depth*? Here the plosive [p] is found in the coda before an affricate, a sound class of higher sonority. The picture should be clear. The English syllable is far more complex than what is suggested in the ASR literature [2], [6], [15].

One of the *a priori* arguments supporting SCH is that since there are fewer syllables than words, a language model composed of syllables describes a reduced search space. An obvious response is that if syllables are so great, why not use any old subdivision of a word, phones for example. Since English has less than four dozen phones, depending on the dialect, there are almost certainly fewer phones than syllables. So, using phones in the language model would reduce the search space even more than would a syllable search space. We contend, however, that the syllable retains linguistic information not present in the phone. Consider the sounds [a] and [t] again. Together they form a distinct syllable in *atlantic*, but not *atrocious*. The distinction between *atrocious* and *atlantic*, expressed as a difference in syllabification, might contribute to more accurate recognition when the phones [a] and [t] are in play. Further, and perhaps related, some researchers have argued that prosodic information, usually not explicitly handled in ASR, is encoded in syllables. Implicitly retaining prosodic information—as would be the case when using syllables rather than phones in the language model—might contribute to more accurate recognition [2].

Finally, it is important to recognize that any syllabifier embodies a theory of the syllable. It is ironic that the syllabifier most frequently cited in the literature and the one used to generate our results is solidly in the tradition of generative linguistics. As we have argued elsewhere, our research is motivated by work in functional, usage-based, and cognitive linguistics [3]. A probabilistic syllabifier would be more appropriate. The ready availability of the NIST syllabifier has dictated our choice however [10]. We defer building a probabilistic syllabifier until later phases of research, but note in passing that such syllabifiers have been developed [8].

Researchers have long-noted the attractiveness of syllables for the acoustic model. Most words are monosyllabic and most syllables are realized acoustically. Yet recognizers with syllables in the acoustic model have shown only modest improvements over conventional triphone-based acoustic models [2]. Syllables are also attractive in the language model for many of the same reasons, as well as for those noted above. Nevertheless, because the output of a recognizer equipped with a syllable language model is a syllable string, published work on syllables in the language model is limited to systems where word strings are unnecessary, systems like children's reading trackers, audio indexing systems, and spoken name recognizers [2]. Since our system does not end with a syllable string, but rather maps these to concept strings, that limitation does not apply.

III. CONCEPTS

The full complexity of human language is redundant in many situations. An air-travel reservation system is an example. In this context certain words and word strings (*fly, flying, going to fly, flew, go to, travelling to, book a ticket to)* mean the same thing. We could express them—and others—in the concept GO. Having collapsed morphology and auxiliary words used to indicate tense, person, aspect, mood, and grouped the base word along with certain formulaic phrases, we have an equivalence of class of words and words strings that mean the same thing. There are certainly fewer concepts, thus defined, than words in English. To take a simple example, *I want to fly to Spokane*, is syllabified as

ay w_aa_n_td t_uw f_l_ay t_uw s_p_ow k_ae_n

and conceptized as

WANT GO s_p_ow k_ae_n.

For Phase I, we hand-generated concepts. Clearly, this will not scale up. In Phase III we will machine-generate concepts using a boot-strapping procedure developed for word-sense disambiguation [3].

Other researchers—at Bell labs, Colorado, and Microsoft Research—have recognized the usefulness of concepts in speech recognition since the early nineties [3]. Our system appears to be novel in that it does not use words, uses probabilistically generated concepts, and is more general than, for example, the utterance classification system developed at Microsoft. Further, taking what appears to be a novel approach, it couples the use of syllables in the language model with probabilistically generated and mapped concepts.

IV. ASR

Since this is not specialist conferences dedicated to ASR, a few words about probabilistic speech recognition seem appropriate. The basic architecture of a recognizer is shown in Figure 2. Its goal is to answer this question: "What is the most likely string of words, *W*, from a language, *L*, given some acoustic input, *A*." Formally:

$$hyp(W) = \frac{argmax}{w \in L} P(W|A) \quad (1)$$

Bayes Rule lets us rewrite equation (1) as:

$$hyp(W) = \frac{argmax}{s \in L} \frac{P(A|W) * P(W)}{P(A)} \quad (2)$$

Since the acoustic signal is the same for all candidate word sequences, (2) may be rewritten as:

$$hyp(W) = \frac{argmax}{s \in L} P(A|W) * P(W) \quad (3)$$

P(A|W) is the acoustic model in Fig. 2. P(W) is the language model. The acoustic model computes conditional probability of a string of words given an acoustic signal in a training corpus. The language model computes the probability that a sequence of words is found in a training corpus. The argmax function lets us pick out the most likely word string. In Fig. 2 it is the decoder, in practice the Viterbi algorithm, an example from dynamic programming. The first two phases of our work substitute syllable strings for word strings in the language model.

Since we are recognizing syllable strings rather than word strings, (3) becomes:

$$hyp(S) = \frac{argmax}{s \in L} P(A|S) * P(S) \quad (4)$$

Phase III introduces a probabilistic concept model that maps concepts to syllable strings, themselves produced by a recognizer with a syllable language model. This is expressed in (5):

$$hyp(C) = \frac{argmax}{C \in M} P(S|C) * P(C) \quad (5)$$

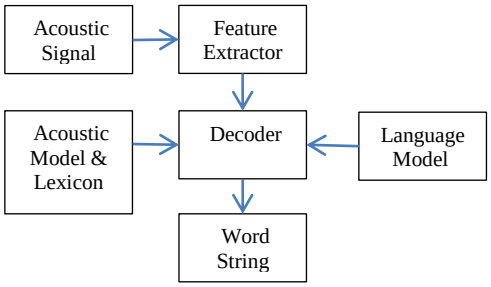

Fig. 2. Basic LVCSR Architecture

V. RESULTS: PHASE I

Results from Phase I have been previously reported [2], [3]. In a nutshell they are:

- Sharp perplexity reduction. Informally, perplexity can be seen as the weighted average branching factor of a language model. Intuitively, it predicts that a recognizer equipped with a syllable language model will perform more accurately than a word language model. We achieved a 37.8% reduction in perplexity with quadrigrams on a small, locally-produced corpus and 85.7% reduction in perplexity on the somewhat larger ATIS0 corpus.
- Substantial recognition improvement. We have achieved a 14.6% mean improvement in recognition accuracy using syllable bigram,

trigram, and quadgrams over the two small corpora mentioned above.

- Slight increase in error rate with concepts. We hand-generated concepts and created a rigid mapping scheme. This produced what we regard as an upper bound error rate increase over a standard recognizer of 1.175%. We hypothesize that recognition accuracy will increase over a standard recognizer once we introduce probabilistically generated concepts and a probabilistic concept mapper [2], [3].

VI. CURRENT WORK

We are currently in Phase II now, reproducing our Phase I results using TIMIT and ATIS2. Phase III, the development of a probabilistic concept model and concept generator is proceeding concurrently. It is interesting to note that our probabilistic concept model closely resembles a probabilistic part-of-speech tagger. The proposed system is shown in Fig. 3. Finally, we must address how to judge the accuracy of the system from input utterance to the output of the concept model. Since the concept strings are human readable, Amazon Mechanical Turk[1] workers will be presented with the initial utterance and the system's output, from both our system and from a conventional recognizer. They will be asked to judge accuracy based on an adaptation of the Likert scale.

Our results to date suggest that using syllables and concepts in automatic speech recognition will improve performance over conventional systems. At minimum, the techniques described here could be adapted to, and improve the performance of, all dialog systems. The usefulness of SCH, however, extends beyond dialog systems to the full range of uses to which automatic speech recognition might be put: freeing digital devices from manual input, increasing the availability of computing to people with sight or dexterity difficulties, and, much, much more. Automatic speech recognition has made enormous strides since the introduction of probabilistic methods two decades ago. We contend that further progress will be made as we attempt to more completely emulate the only know fully functional speech processor, the human being.

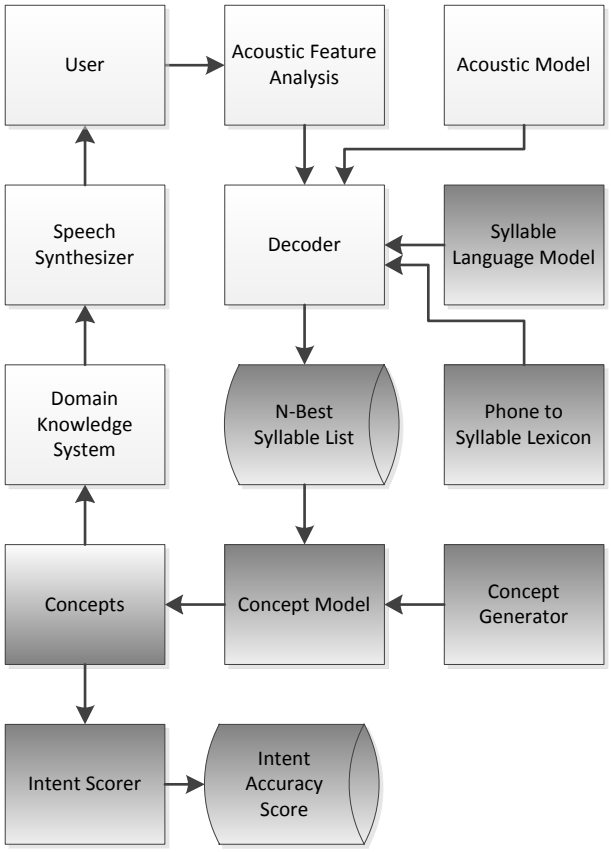

Fig. 3. Acoustic features are decoded into syllable strings using a syllable language model. The syllables strings are probabilistically mapped to concept strings. The N-best syllable list is rescored using concepts. The Intent Scorer enables comparison of performance with a conventional recognizer.

[1] https://www.mturk.com/mturk/welcome

REFERENCES[2]

[1] N. Cowan, T. Keller, C. Hulme, S. Roodenkys, S. McDougal, S., J. Rack, J. 1994, "Verbal memory span in humans: speech timing clues to the mechanisms underlying age and word length effects," Journal of Memory and Language, vol. 33, pp. 234-250, 1994.

[2] P. De Palma, Syllables and Concepts in Large Vocabulary Speech Recognition, Ph.D. Dissertation, Department of Linguistics, University of New Mexico, Albuquerque, NM, 2010.

[3] P. De Palma, G. Luger, C. Smith, W. Wooters, "Bypassing words in automatic speech recognition, Proceedings of the 23rd Midwest Artificial Intelligence and Cognitive Science Conference, Cincinnati, OH, April 21, 2012.

[4] W. Foley, William, "Field methods," in The Linguistics Encylopedia, K. Malmkjaer, Ed. London: Routledge, 1999.

[5] D. Jurafsky, J. Martin, Speech and Language Processing, Upper Saddle River, NJ: Pearson/Prentice Hall, 2009.

[6] M. Kenstowicz, Phonology in Generative Grammar. Oxford: Blackwells, 1999.

[7] K. Kohler ,"Is the syllable a phonological universal?", Journal of Linguistics, vol. pp. 207-208, 1966.

[8] Y. Marchand, C. Adsett, R. Damper, "Evaluating Automatic Syllabification Algorithms for English," Proceedings of the 6th International Conference of the Speech Communication Association, pp. 316-321, 2007.

[9] J. Miller, R. Weinert, Spontaneous Spoken Language, Oxford: Oxford U. Press, 1998, p. 1.

[10] NIST, "Language Technology Tolls/Multimodel Information Group—Tools," Retrieved Feb. 19, 2012 from: http://www.nist.gov.

[11] NIST-RT, "Rich Transcription Evaluation Project. National Institute of Standards, Information Access Division." Retrieved May 23, 2008 from: http://www.nist.gov/speech/tests/rt/.

[12] L. Rabiner, B. Juang, B., Fundamentals of Speech Recognition. Englewood Cliffs, NJ: Prentice Hall, 1993.

[13] C. Sanh-Hun, "In Myanamar, fear is ever present," The New York Times, October 21, 2007.

[14] C. Schmidt-Nielsen, T. Crystal, T., "Speaker Verification by Human Listeners: Experiments Comparing Human and Machine Performance Using NIST 1998 Speaker Evaluation Data," Digital Signal Processing, vol. 10, 249-266, 2000.

[15] E. Selkirk, "The syllable," in Phonological Theory: The Essential Readings, J. Goldsmith, Ed. Malden, MA: Blackwell, 1999, pp. 328-350.

[16] M. Tomasello, The New Psychology of Language, Vol. 2. Lawrence Erlbaum and Associates.

[2] For the sake of brevity, only items either quoted, immediately relevant, or not easily available through a on-line search are included here. A complete set of references may be found in [2] and [3].