Bypassing Words in Automatic Speech Recognition

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

George Luger
Department of Computer Science
University of New Mexico
Albuquerque, NM
luger@cs.unm.edu

Caroline Smith
Department of Linguistics
University of New Mexico
Albuquerque, NM
caroline@unm.edu

Charles Wooters
Next It Corporation
Spokane, WA
cwooters@nextit.com

Abstract

Automatic speech recognition (ASR) is usually defined as the transformation of an acoustic signal to words. Though there are cases where the transformation to words is useful, the definition does not exhaust all contexts in which ASR could be used. Once the constraint that an ASR system outputs words is relaxed, modifications that reduce the search space become possible: 1) The use of syllables instead of words in the recognizer's language model; 2) The addition of a concept model that transforms syllable strings to concept strings, where a concept collects related words and phrases. The paper presents preliminary positive results on the use of syllables and concepts in speech recognition and outlines our current efforts to verify the Syllable-Concept Hypothesis (SCH).

Introduction

The speech recognition problem is conventionally formulated as the transformation of an acoustic speech signal to word strings. Yet this formulation dramatically underspecifies what counts as word strings. Here is a "33-year-old business woman" speaking to a reporter from *The New York Time***s**: "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" (Sang-Hun 2007: 1). The reporter has certainly transformed an acoustic signal into words. Though it would be nice to have a recording and transcription of the actual interview, we can get a sense of what the reporter left out (and put in) by looking at any hand-transcribed corpus of spontaneous speech. Here is the very first segment from 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.

While we recognize the benefits of solving the speech recognition problem as described, the research presented here begins with the observation that human language performance does not include transcription from an acoustic signal to words—either in the sanitized form found in The *New York Times* quote or in the raw form found in the Buckeye Corpus. We do not suggest that AI research limit itself to human performance. We do claim that there is much to be gained by relaxing the constraint that the output of automatic speech recognition be a word string. Consider a speech recognition system designed to handle spoken plane reservations via telephone or, for that matter, just about any spoken-dialog system. The recognizer need only pass on the sense of the caller's speech to an appropriately constructed domain knowledge system to solve a problem of significant scope.

The question of what is meant by the sense of an utterance is central to this research. As a first approximation, one can think of the sense of an utterance as a sequence of concepts, where a concept is an equivalence class of words and phrases that seem to mean the same thing. A conventional recognizer generates a word string given a sequence of acoustic observations. The first stage in our research is to generate a syllable string given the same sequence of acoustic observations. Notice that the search space is much reduced. There are

fewer syllables to search through (and mistake) than words. Of course, this syllable string must undergo a further transformation to be useful. One possibility would be to probabilistically map it to word strings. We have experimented with this. The results have not been encouraging. We propose, instead, to generate a concept string given the syllable string. Once again, the search space is reduced. There are fewer concepts to search through and mistake than words.

The Symbol-Concept Hypothesis (SCH) claims that this dual reduction in search space will result in better recognition accuracy over a standard recognizer. Though SCH can be argued using the axioms of probability, at bottom it is an empirical hypothesis. Preliminary experimental results have been promising. This paper is the first in a four phase, multi-year research effort to test SCH:

- Phase I: Gather preliminary data about SCH using small corpora.
- Phase II: Reproduce the results from Phase I using a much larger corpus.
- Phase III: Introduce a probabilistic concept generator and concept model.
- Phase IV: Introduce an existing domain knowledge system and speech synthesizer to provide response to the user.

Background

The goal of probabilistic speech recognition is to answer this question: "What is the most likely string of words, *W*, from a language, *L*, given some acoustic input, *A*." This is formulated in equation 1:

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

Since words have no place in SCH, we speak instead of symbol strings drawn from some set of legal symbols, with the sole constraint that the symbols be encoded in ASCII format. So, equation (1) becomes:

$$hyp(S) = \frac{argmax}{s \in L} P(S|A) \qquad (2)$$

Equation (2) is read: "The hypothesized symbol string is the one with the greatest probability given the sequence of acoustic observations" (De Palma 2010:16). Bayes Theorem lets us rewrite equation (2) as:

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

Since *P(A)* does not affect the computation of the most probable symbol string (the acoustic observation is the acoustic observation, no matter the potential string of symbols) we arrive at a variation of the standard formulation of probabilistic speech recognition (Jurafsky and Martin 2009):

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

The difference is that the formulation has been generalized from words to any symbol string. *P(A|S)*, known as the likelihood probability in Bayesian inference, is called the acoustic model in the context of automatic speech recognition. *P(S)*, known as the prior probability in Bayesian inference, is called the language model in ASR. The acoustic model expresses the probability that a string of symbols—words, syllables, whatever—is associated with an acoustic signal in a training corpus. The language model expresses the probability that a sequence of symbols—again, words, syllables, whatever—is found in a training corpus.

The attractiveness of syllables for the *acoustic model* of speech recognition has been noted for some time. A study of the SWITCHBOARD corpus found that over 20% of the manually annotated phones are never realized acoustically, since phone deletion is common in fluent speech. On the other hand, the same study showed that 99% of canonical syllables are realized in speech. Syllables also have attractive distributional properties. The statistical distributions of the 300 most frequently occurring words in English and the most common syllables are almost identical. Though monosyllabic words account for only 22% of SWITCHBOARD by type, they account for a full 81% of tokens (Greenberg 1999; Greenberg 2001; Greenberg et al. 2002). All of this suggests that the use of syllables in the acoustic model might avoid some of the difficulties associated with word pronunciation variation due to dialect, idiolect, speaking rate, acoustic environment, and pragmatic/semantic context.

Nevertheless, most studies indicate positive but not dramatic improvement when using a syllable-based acoustic model (Ganapathiraju et al. 1997 and 2002; Sethy and Narayanan 2003; Hamalainen et al. 2007). This has been disappointing given the theoretical attractiveness of syllables in the acoustic model. Since this paper is concerned exclusively with the language model and post-language model processing, conjectures about the performance of syllables in the acoustic model performance are beyond its scope.

Still, many of the reasons that make syllables attractive in the acoustic model also make them attractive in the language model, including another not mentioned in the literature on acoustic model research: there are fewer syllables than words, a topic explored later in this paper. Since the output of a recognizer using a syllable language model is a syllable string, studies of speech recognition using syllable language models have been limited to special purpose systems where output word strings are not necessary. These include reading trackers, audio indexing

systems, and spoken name recognizers. Investigations report significant improvement over word language models (Bolanos et al. 2007; Schrumpf, Larson, and Eickler 2005; Sethy and Narayanan 1998). The system proposed here, however, does not end with a syllable string, but, rather, passes this output to a concept model—and thereby transforms them to concept strings, all to be described later.

Researchers have recognized the potential usefulness of concepts in speech recognition: since the early nineties at Bell Labs, later at the University of Colorado, and still later at Microsoft Research (Pieraccini et al. 1991; Hacioglu and Ward 2001; Yaman et al. 2008). The system proposed here does not use words in any fashion (unlike the Bell Labs system), proposes the use of probabilistically generated concepts (unlike the Colorado system), and is more general than the utterance classification system developed at Microsoft. Further, it couples the use of sub-word units in the language model, specifically syllables, with concepts, an approach that appears to be novel.

Syllables, Perplexity, and Error Rate

One of the first things that a linguist might notice in the literature on the use of the syllable in the acoustic model is that its complexity is underappreciated. Rabiner and Juang (1993), an early text on speech recognition, has only two index entries for "syllable" and treat it as just another easily-defined sub-word unit. This is peculiar, since the number of English syllables varies by a factor of 30 depending on whom one reads (Rabiner and Juang 1993; Ganapathiraju, et al. 1997; Huang et al. 2001). In fact, there is a substantial linguistic literature on the syllable and how to define it across languages. This is important since any piece of software that claims to syllabify words embodies a theory of the syllable. Thus, the syllabifier that is cited most frequently in the speech recognition literature, and the one used in the work described in this paper, implements a dissertation that is firmly in the tradition of generative linguistics (Kahn 1976). Since our work is motivated by more recent research in functional and cognitive linguistics (see, for example, Tomasello 2003), a probabilistic syllabifier might be more appropriate. We defer that to a later stage of the project, but note in passing that probabilistic syllabifiers have been developed (Marchand, et al. 2007).

Still, even though researchers disagree on the number of syllables in English, that number is significantly smaller than the number of words. And therein lies part of their attractiveness for this research. Simply put, the syllable search space is significantly smaller than the word search space. Suppose language *A* has *a* words and language *B* has *b* words, where $a > b$. All other things being equal, the probability of correctly guessing a word from *B* is greater than guessing one from *A*. Suppose further, that these words are not useful in and of themselves, but contribute to some downstream task, the accuracy of which is proportional to the accuracy of the word recognition task. Substitute syllables for words in language B—since both are symbols—and this is exactly the argument being made here.

Now, one might ask, if syllables work so nicely in the language model of speech recognition, why not use another sub-word with an even smaller symbol set, say a phone or demi-syllable? Though the question is certainly worth investigating empirically, the proposed project uses syllables because they represent a compromise between a full word and a sound. By virtue of their length, they preserve more linguistic information than a phone and, unlike words they represent a relatively closed set. Syllables tend not to change much over time.

A standard *a priori* indicator of the probable success of a language model is lower perplexity, where perplexity is defined as the N^{th} inverse root of the probability of a sequence of words (Jurafsky and Martin 2009; Ueberla 1994):

$$PP(W) = p(w_1 w_2 \ldots w_n)^{-1/n} \qquad (5)$$

Because there are fewer syllables than words, we would expect both their perplexity in a language model to be lower and their recognition accuracy to be higher. Since the history of science is littered with explanations whose self-evidence turned out to have been incorrect upon examination, we offer a first pass at an empirical investigation.

To compare the perplexity of both syllable and word language models, we used two corpora, the Air Travel Information System (Hemphill 1993) and a smaller corpus (SC) of human-computer dialogs captured using the Wizard-of-Oz protocol at Next It (Next IT 2012), where subjects thought they we were interacting with a computer but in fact were conversing with a human being. The corpora were syllabified using software available from the National Institute of Standards and Technology (NIST 2012).

Test and training sets were created from the same collection of utterances, with the fraction of the collection used in the test set as a parameter. The results reported here use a randomly chosen 10% of the collection in the test set and the remaining 90% in the training set. The system computed the mean, median, and standard deviation over twenty runs. These computations were done for both word and syllable language models for unigrams, bigrams, trigrams, and quadrigrams (sequences of one, two, three, and four words or syllables). As a baseline, the perplexity of the unweighted language model—one in which any word/syllable has the same probability as any other—was computed.

For bigrams, trigrams, and quadrigrams, the perplexity of a syllable language model was less than that of a word language model. Of course, in comparing the perplexity of syllable and word language models, we are comparing

sample spaces of different sizes. This can introduce error based on the way perplexity computations assign probability mass to out-of-vocabulary tokens. It must be recognized, however, that syllable and word language models are not simply language models of different sizes of the kind that Ueberla (1994) considered. Rather, they are functionally related to one another. This suggests that the well-understood caution against comparing the perplexity of language models with different vocabularies might not apply completely in the case of syllables and words. Nevertheless, the drop in perplexity was so substantial in a few cases (37.8% SC quadrigrams, 85.7% ATIS bigrams), that it invited empirical investigation with audio data.

Recognition Accuracy

Symbol Error Rate (SER) is the familiar Word Error Rate (WER) generalized so that context clarifies whether we are talking about syllables, words, or concepts. The use of SER raises a potential problem. The number of syllables (either by type or token) differs from the number of words in the training corpus. Further, in all but monosyllabic training corpora, syllables will, on average, be shorter than words. How then can we compare error rates? The answer, as before, is that 1) words are functionally related to syllables and 2) improved accuracy in syllable recognition will contribute to downstream accuracy in concept recognition.

To test the hypothesis that a syllable language model would perform more accurately than a word language model, we gathered eighteen short audio recordings, evenly distributed by gender, and recorded over both the public switched telephone network and mobile phones. The recognizer used was SONIC from the Center for Spoken Language Research of the University of Colorado (SONIC 2010). The acoustic model was trained on the MACROPHONE corpus (Bernstein et al. 1994). Additional tools included a syllabifier and scoring software available from the National Institute of Standards and Technology (NIST 2012), and language modeling software developed by one of the authors.

The word-level transcripts in the training corpora were transformed to phone sequences via a dictionary look-up. The phone-level transcripts were then syllabified using the NIST syllabifier. The pronunciation lexicon, a mapping of words to phone sequences, was similarly transformed to map syllables to phone sequences. The word-level reference files against which the recognizer's hypotheses were scored were also run through the same process as the training transcripts to produce syllable-level reference files.

With these alterations, the recognizer transformed acoustic input into syllable output represented as a flavor of Arpabet. Figure 1 shows an extract from a reference file represented both in word and in phone-based syllable form.

i want to fly from spokane to seattle
ay waantd tuw flay frahm spow kaen tuw si ae dxaxl

i would like to fly from seattle to san Francisco
ay wuhdd laykd tuw flay frahm siy ae dxaxl tuw saen fraen sih skow

Figure 1: Word and Syllable References

The recognizer equipped with a syllable language model showed a mean improvement in SER over all N-gram sizes of 14.6% when compared to one equipped with a word language model. Though the results are preliminary, and await confirmation with other corpora, and with the caveats already noted, they suggest that a recognizer equipped with a syllable language model will perform more accurately than one equipped with a word language model.[1] This will contribute to the downstream accuracy of the system described below. Of course, it must be pointed out that some of this extraordinary gain in recognition accuracy will necessarily be lost in the probabilistic transformation to concept strings.

Concepts

At this point one might wonder about the usefulness of syllable strings, no matter how accurately they are recognized. We observe that the full range of a natural language is redundant in certain pre-specified domains, say a travel reservation system. Thus the words and phrases *ticket, to book a flight, to book a ticket, to book some travel, to buy a ticket, to buy an airline ticket, to depart, to fly, to get to*, all taken from the reference files for the audio used in this study, describe what someone wants in this constrained context, namely to go somewhere. With respect to a single word, we collapse morphology and auxiliary words used to denote person, tense, aspect, and mood, into a base word. So, *fly, flying, going to fly, flew, go to, travelling to*, are grouped, along with certain formulaic phrases (*book a ticket to*), in the equivalence class, GO. Similarly, the equivalence class WANT contains the elements *buy, can I, can I have, could I, could I get, I like, I need, I wanna, I want, I would like, I'd like, I'd like to have, I'm planning on, looking for, need, wanna, want, we need, we would like, we'd like, we'll need, would like.* We refer to these equivalence classes as concepts.

For example, a sentence from the language model (I want to fly to Spokane) was syllabified, giving:

ay w_aa_n_td t_uw f_l_ay t_uw s_p_ow k_ae_n

[1] Though it might be interesting and useful to look at individual errors, the point to keep in mind is that we are looking for broad improvement. The components of SCH were not so much arguments as the initial justification for empirical investigations, investigations that will support or falsify SCH.

Then concepts were mapped to the syllable strings, producing:

WANT GO s_p_ow k_ae_n

The mapping from concepts to syllable strings was rigid and chosen in order to generate base-line results. The mapping rules required that at least one member of an equivalence class of syllable strings had to appear in the output string for the equivalence class name to be inserted in its place in the output file. For example, k_ae_n ay hh_ae_v (*can I have*) had to appear in its entirety in the output file for it to be replaced with the concept WANT.

The experiment required that we:

1. Develop concepts/equivalence classes from the training transcript used in the language model experiments.
2. Map the equivalence classes onto the reference files used to score the output of the recognizer. For each distinct syllable string that appears in one of the concept/equivalence classes, we substituted the name of the equivalence class for the syllable string. We did this for each of the 18 reference files that correspond to each of the 18 audio files. For example, WANT is substituted for every occurrence of *ay w_uh_dd l_ay_kd* (*I would like*).
3. Map the equivalence classes onto the output of the recognizer when using a syllable language model for N-gram sizes 1 through 4. We mapped the equivalence class names onto the content of each of the 72 output files (4 x 18) generated by the recognizer.
4. Determine the error rate of the output in step 3 with respect to the reference files in step 2.

As before, the SONIC recognizer, the NIST syllabifier and scoring software, and our own language modeling software were used. The experiments showed a mean increase in SER over all N-gram sizes of just 1.175%. Given the rigid mapping scheme, these results were promising enough to encourage us to begin work on: 1) reproducing the results on the much larger ATIS2 corpus (Garofalo 1993) and 2) a *probabilistic* concept model.

Current Work

We are currently building the system illustrated in Figure 2. The shaded portions describe our work. A crucial component is the concept generator. Under our definition, concepts are purely collocations of words and phrases, effectively, equivalence classes. In order for the system to be useful for multiple domains, we must go beyond our preliminary investigations: the concepts must be machine-generated. This will be done using a boot-strapping procedure, first described for word-sense disambiguation. The algorithm takes advantage of "the strong tendency of words to exhibit only one sense per collocation and per discourse" (Yarowsky 1995: 50). The technique will begin with a hand-tagged seed set of concepts. These will be used to incrementally train a classifier to augment the seed concepts. The output of a speech recognizer equipped with a syllable language model is the most probable sequence of syllables given an acoustic event. The formalisms used to probabilistically map concepts to syllable strings are reworkings of equations (1) to (4), resulting in:

$$hyp(C) = \frac{argmax}{c \in M} P(C|S) = \frac{argmax}{C \in M} P(S|C) * P(C) \quad (6)$$

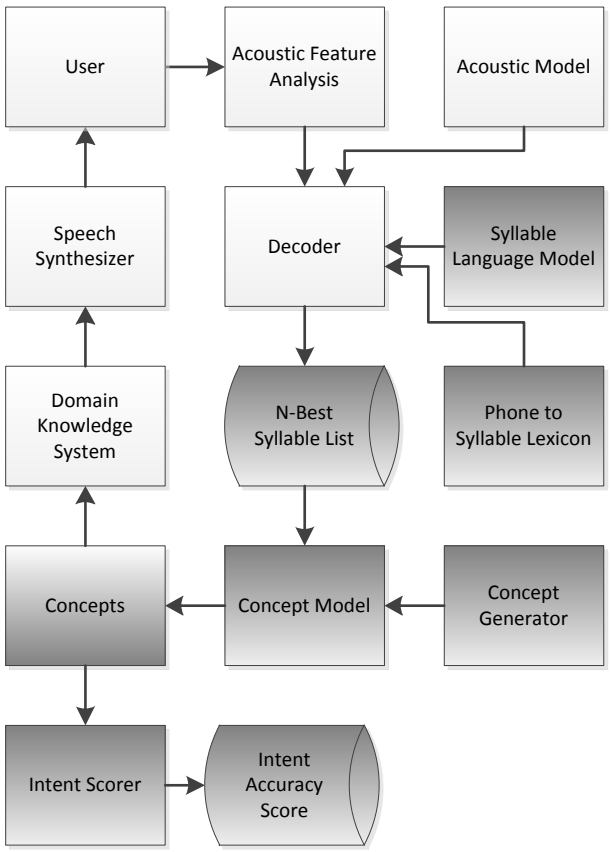

Figure 2: Acoustic features are decoded into syllable strings using a syllable language model. The syllables strings are probabilistically mapped to concept strings. The N-best syllable list is rescored using concepts. The Intent Scorer enables comparison of performance with a conventional recognizer.

M is just the set of legal concepts created for a domain by the concept generator. Equation (6) is an extension of a classic problem in computational linguistics: probabilistic part-of-speech tagging. That is, given a string of words, what is the most probable string of parts-of-speech? In the case at hand, given a syllable string, what is the most probable concept string?

Using equation (6), the Syllable-Concept Hypothesis, introduced early in the paper, can be formalized. If equation (1) describes how a recognizer goes about choosing a word string given a string of acoustic observations, then our enhanced recognizer can be described in equation (7):

$$hyp(C) = \frac{argmax}{C \in M} P(C|A) \quad (7)$$

That is, we are looking for the legal concept string with the greatest probability given a sequence of acoustic observations. SCH, in effect, argues that the P(C|A) exceeds the P(W|A).

Finally, the question of how to judge the accuracy of the system, from the initial utterance to the output of the concept model, must be addressed. Notice that the concept strings themselves are human readable. So,

I WANT TO FLY TO SPOKANE

becomes:

WANT GO s_p_ow k_ae_n

Amazon Mechanical Turk[2] workers will be presented with both the initial utterance as text and the output of the concept model as text and asked to offer an opinion about accuracy based on an adaptation of the Likert scale. To judge how the proposed system performs relative to a conventional recognizer, the same test will be made, substituting the output of the recognizer with a word language model and no concept model for the output of the proposed system.

Conclusion

We have argued that the speech recognition problem as conventionally formulated—the transformation of an acoustic signal to words—neither emulates human performance nor exhausts the uses to which ASR might be put. This suggests that we could bypass words in some ASR applications, going from an acoustic to signal to probabilistically generated syllable strings and from there to probabilistically generated concept strings. Our experiments with syllables on small corpora have been promising:

- 37.8% drop in perplexity with quadrigrams on the SC corpus
- 85.7% drop in perplexity with ATIS bigrams
- 14.6% mean increase in recognition accuracy over bigram, trigram, and quadrigrams

But as has been pointed out, a syllable string is not useful in a dialog system. Concepts must be mapped to syllables. A concept, as we define it, is an equivalence class of words and phrases that seem to mean the same thing in a given context. To date, we have hand-generated concepts from reference files and mapped them to syllables using a ridgid mapping scheme intended as a baseline.

But to be truly useful, any recognizer using concepts must automatically generate them. Since concepts, under our definition, are no more than collocations of words, we we propose a technique first developed for word-sense disambiguation: incrementally generate a collection of concepts from a hand-generated set of seed concepts. The idea in both phases or our work—probabilistically generating syllable strings and probabilistically generating concept strings—is to reduce the search space from what conventional recognizers encounter. At the very end of this process, we propose scoring how closely the generated concepts match the intent of the speaker using Mechanical Turk workers and a modified Likert scale. Ultimately the output the system will be sent on to a domain knowledge system, from there onto a speech synthesizer, and finally to the user, who, having heard the output will respond, thus starting the cycle over gain.

Our results to date suggests that the use of syllables and concepts in ASR will results in improved recognition accuracy over a conventional word-based speech recognizer. This improved accuracy has the potential to be used in fully functional dialog systems. The impact of such systems could be as far-reaching as the invention of the mouse and windowing software, opening up computing to persons with coordination difficulties or sight impairment, freeing digital devices from manual input, and transforming the structure of call centers. One application, often overlooked in catalogues of the uses to which ASR might be put, is surveillance.[3] The Defense Advanced Research Agency (DARPA) helped give ASR its current shape. According to some observers, the NSA, as a metonym for all intelligence agencies, is drowning in unprocessed data, much of which is almost certainly speech (Bamford 2008). The kinds of improvements described in this paper, the kinds that promise to go beyond the merely incremental, are what are needed to take voice recognition to the next step.

References

Bamford, J. 2008. *The Shadow Factory: The Ultra-Secret NSA from 9/11 to the Eavesdropping on America*. NY: Random House.

Bernstein, J., Taussig, K., Godfrey, J. 1994.

[2] The Amazon Mechanical Turk allows computational linguists (and just about anyone else who needs a task that requires human intelligence) to crowd-source their data for human judgment. See https://www.mturk.com/mturk/welcome

[3] Please note that this paper is not necessarily an endorsement of all uses to which ASR might be put. It merely recognizes what is in fact the case.

MACROPHONE. Linguistics Data Consortium, Philadelphia PA

Bolanos, B., Ward, W., Van Vuuren, S., Garrido, J. 2007. Syllable Lattices as a Basis for a Children's Speech Reading Tracker. *Proceedings of Interspeech-2007*, 198-201.

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

Ganapathiraju, A., Goel, V., Picone, J., Corrada, A., Doddington, G., Kirchof, K., Ordowski, M., Wheatley, B. 1997. Syllable—A Promising Recognition Unit for LVCSR. *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 207-214.

Ganapathiraju, A., Hamaker, J., Picone, J., Ordowski, M., Doddington, G. 2001. Syllable-based large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, 358-366.

Garofalo, J. 1993. ATIS2. Linguistics Data Consortium, Philadelphia, PA

Greenberg, S. 1999. Speaking in Shorthand—A Syllable-Centric Perspective for Understanding Pronunciation Variation. *Speech Communication*, 29, 159-176.

Greenberg, S. 2001. From here to Utility—Melding Insight with Speech Technology. *Proceedings of the 7th European Conference on Speech Communication and Technology*, 2485-2488.

Greenberg, S., Carvey, H. Hitchcock, L., Chang, S. 2002. Beyond the Phoneme: A Juncture-Accent Model of Spoken Language. *Proceedings of the 2nd International Conference on Human Language Technology Research*, 36-43.

Hacioglu, K., Ward, W. 2001. Dialog-Context Dependent Language Modeling Combining N-Grams and Stochastic Context-Free Grammars. *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 537-540.

Hamalainen, A., Boves, L., de Veth, J., Bosch, L. 2007. On the Utility of Syllable-Based Acoustic Models for Pronunciation Variation Modeling. *EURASIP Journal on Audio, Speech, and Music Processing*, 46460, 1-11.

Hemphill, C. 1993. ATIS0. Linguistics Data Consortium, Philadelphia, PA.

Huang, X., Acero, A., Hsiao-Wuen, H. 2001. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, NJ: Prentice Hall.

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

Kahn, D. 1976. *Syllable-based Generalizations in English Phonology*. Ph.D. dissertation, Department of Linguistics, University of Indiana, Bloomington, In: Indiana University Linguistics Club.

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

Next It Corporation. 2012. Web Customer Service with Intelligent Virtual Agents. Retrieved 3/37/2012 from: http://www.nextit.com.

NIST. 2012. Language Technology Tools/Multimodel Information Group—Tools. Retrieved 2/19/2012 from: http://www.nist.gov.

Pieraccini, R., Levin, E., Lee, C., 1991. Stochastic Representation of Conceptual Structure in the ATIS Task. *Proceedings of the DARPA Speech and Natural Language Workshop, 121-124*.

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

Sang-Hun, C. 10/21/2007. Myanmar, Fear Is Ever Present. *The New York Times.*

Schrumpf, C., Larson, M., Eickler, S., 2005. Syllable-based Language Models in Speech Recognition for English Spoken Document Retrieval. *Proceedings of the 7th International Workshop of the EU Network of Excellence DELOS on Audio-Visual Content and Information Visualization in Digital Libraries*, pp. 196-205.

Sethy, A., Narayanan, S. 2003. Split-Lexicon Based Hierarchical Recognition of Speech Using Syllable and World Level Acoustic Units, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, I, 772-775.

SONIC. 2010. SONIC: Large Vocabulary Continuous Speech Technology. Retrieved 3/8/2010 from: http://techexplorer.ucsys.edu/show_NCSum.cfm?NCS=258626.

Tomasello, M. (ed.) 2003. *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ueberla, J. 1994. *Analyzing and Improving Statistical Language Models for Speech Recognition*. Ph.D. Dissertation, School of Computing Science, Simon Frazier University.

Yaman, S., Deng, L., Yu, D., Wang, W, Acera, A. 2008. An Integrative and Discriminative Technique for Spoken Utterance Classification. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, 1207-1214.

Yarowsky, D. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 189-196.