

Phonetic Techniques for Achieving High Accuracy in Spoken Access to Very Large Lists

Melvyn J. Hunt¹ & Yoon Kim²

Novauris Technologies Ltd

¹Millbank
Stoke Road
Bishops Cleeve
Cheltenham
GL52 8RW
England

melvyn.hunt@novauris.com

²1590 Drew Ave.
Suite 100
Davis
CA 95616
USA

yoon.kim@novauris.com

Summary

This paper sets out the Novauris view that incorporation of phonetic knowledge in automatic speech recognition systems can make an important contribution to recognition accuracy, provided that the knowledge is applied in an objective, usually quantitative, way and not imposed as a set of subjectively derived rules. In addition to handling very large lists for spoken access (*e.g.* US name-and-address recognition), Novauris has needed high accuracy as it has turned its attention to spoken access to list items in which the spoken items are relatively short and thus have less acoustic information than was the case for names and addresses. Some phonetic properties of English syllables are discussed, particularly the differences between prevocalic and postvocalic consonants. To illustrate the language-specificity of these properties, the very different properties of French and Korean are described. Techniques are then described for improving the large Novauris pronouncing dictionary. Finally, some performance figures are provided on speech applications that take advantage of the work described in the paper.

1. Introduction

Novauris specializes in applications of speech recognition technology that provide speaker-independent spoken access to large lists either directly or over the telephone. We combine our speech recognizer (employing more knowledge at the acoustic-phonetic level than conventional systems) with fast symbolic search techniques in a closely integrated manner that results in rapid, accurate, robust access with relatively small computational load. We believe that interactions are faster and more agreeable for the user if the list item can be specified through a single utterance rather than requiring a question-response sequence, possibly including multi-stage confirmations [1].

Our first technology demonstration allowed single-utterance, speaker-independent access to 245 million US-style names and addresses. The addresses were real down to the specification of the street, but the house numbers and the associated names were synthesized from statistics made available by the US Postal Service and the US Census Bureau respectively. Tests with direct microphone input from 50 held-out native speakers of American English under matched recording conditions gave a recognition error rate of only 0.2%. The median response time on a standard PC with a 2 GHz processor was around a third of a second. Less formal tests indicated that the error rate with telephone speech would, not surprisingly, be higher, but it was certainly less than 1%.

Our approach to obtaining accurate spoken access to items in large lists depends to some extent on efficient exploitation of redundant information within the item. This is particularly true of names and addresses, where there is extensive overlap of information, for example between the name (even though names are not unique) and the address and within the address between the Zip code and the state and city.

Since completing the name-and-address work, we have turned our attention to lists containing shorter items, such as those needed for destination entry to in-car navigation systems, for specifying music items from extensive MP3-type playlists or TV programs selected from electronic programming guides. In these cases, the redundancy is much reduced, and the accuracy of our speaker-independent speech recognition itself becomes relatively more important, compared to the back-end symbolic search. Novauris has always believed it important to exploit phonetic knowledge to a greater extent than is usual in most current speech recognition technology, always provided that this knowledge is derived in a sound statistical manner and is not merely heuristics or an expression of human subjective impressions and prejudices. This paper describes some of the phonetic considerations that we believe are important to ensure high accuracy and that are contributing to our ability to offer reliable selection from low-redundancy items in a large database.

2. Syllable Structure in English

Traditionally, large-vocabulary automatic speech recognition systems describe words as sequences of phonemes. Despite some popular misconceptions, phonemes are not acoustic units but rather linguistic units. They may perhaps have a definite psychological existence, but they are not well defined acoustically: some cannot be perceived in isolation, while others merge their identifying cues with neighboring phonemes. To some extent, these phenomena can be handled through context-dependent phonetic units, specifically triphones. However, there are several phenomena that triphones do not typically account for.

One of the key phenomena generally ignored by triphones is the structure of syllables. Consonants coming after the vowel (*postvocalic* consonants) in a syllable behave quite differently in English from those coming before the vowel (*prevocalic* consonants). Figure 1 shows how the error rate and detectability of individual consonants varies widely depending on whether they are prevocalic or postvocalic.

The lexical stress assigned to a syllable in a word, primarily affecting the vowel, is another factor that is often ignored in automatic speech recognition systems. It primarily affects the duration, loudness and pitch contour of the vowel, but unstressed vowels also tend to be more centralized and more variable. A Novauris phonetic recognition system found to have 81% accuracy in recognizing vowels in syllables with primary stress, was found to achieve only 49% for vowels in unstressed syllables.

Stress also has a strong influence on the way certain consonants are produced, such as the tendency of certain consonants to be realized as so-called flaps, which, in the case of /t/, even turns a voiceless consonant into a voiced consonant.

Stress also strongly influences the tendency of prevocalic voiceless plosives (/p/, /t/, /k/ in English) to be aspirated or not. (*Aspiration* is the production of a puff of air similar to /h/ between the release of the voiceless plosive and the onset of the vowel. It does not occur in, for example, French, Spanish or Italian.)

Phonetic recognition results for voiceless plosives in American English are shown in Figures 1, 2 and 3 (derived from results presented in [2]). They show the influence of location within the syllable, prevocalic or postvocalic, and they show the influence of syllable stress. Broadly speaking, plosives are recognized more reliably in stressed syllables (bars 1 and 2) than in unstressed syllables (bars 3 and 4). In stressed syllables, prevocalic plosives (bar 1) are recognized more reliably than postvocalic plosives (bar 2), though in unstressed syllables there is little or no difference between the prevocalic (bar 3) and postvocalic (bar 4) position.

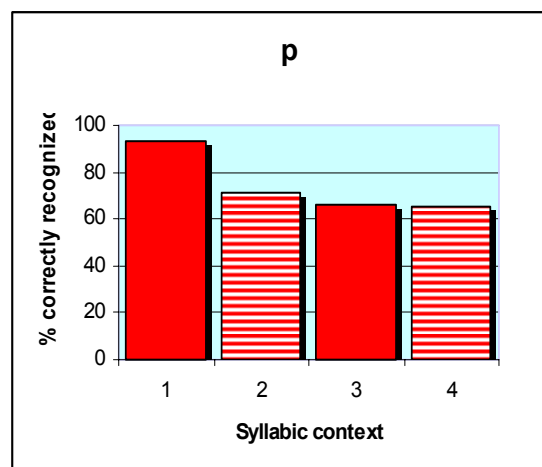


Figure 1: Average recognition accuracy for /p/ in: 1 & 2 stressed syllables; 3 & 4 unstressed syllables; 1 and 3 prevocalic position; 2 & 4 postvocalic position.

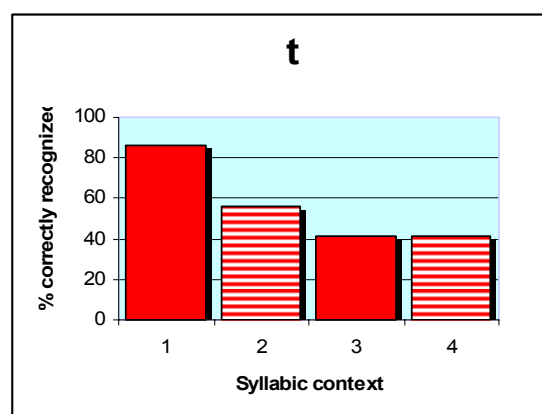


Figure 2: Recognition accuracy for /t/, with bars as in Figure 1.

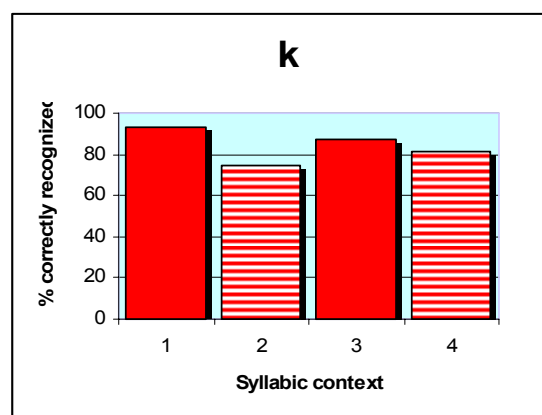


Figure 3: Recognition accuracy for /k/, with bars as in Figure 1.

3. Syllable Structure in Other Languages – Comparison with French and Korean

The allowed range of syllable structures, and the properties that follow from them, depend on the particular language being spoken. Indo-European languages, particularly the Germanic branch, including English, German, Swedish, *etc* and the Slavic branch, including Russian, Polish, *etc*, are unusual among the world's languages in their complex syllable structures and large lexical stress distinctions. English, for example, allows up to three prevocalic consonants (for example in *string*) and up to four postvocalic consonants (for example in *texts*).

The majority of the world's languages allow fewer prevocalic consonants and especially fewer postvocalic consonants, with many preferring so-called open syllables, which have no postvocalic consonants. Among Western languages, Spanish, French, Italian and Portuguese have notably simpler syllable structures with a preference for open syllables, while Asian languages including Chinese, Japanese and Korean are even more extreme in this respect.

A pair of these languages, namely Korean, spoken with native competence by one of the co-authors, and French, spoken with reasonable competence by the other co-author, exemplify ways in which languages can differ from English in their syllabic behavior.

Neither of these languages have strong lexical stress distinctions. Both of them allow a maximum of only two prevocalic consonants¹, and French allows up to only two postvocalic consonants, while Korean allows only one [3]. In these two respects, then, French and Korean are more similar to each other than either is to English. However, in another respect they lie at two extremes with respect to English. This is in the difference between prevocalic and postvocalic realizations of the same consonant phoneme.

Let us look at plosives (in English they are / p, t, k, b, d, g /) first. Prevocalic plosives always have to be *released*: that is, the obstruction caused by the tongue or lips in the consonant has to be opened up to allow the vowel to be produced, and it is the sudden opening that makes most of the sound that we associate with the plosive. In English, postvocalic plosives, on the other hand, do not have to be released; it depends on the phonetic context and on how carefully we are speaking. In fact, despite entreaties from some schoolteachers and some mothers, most of us do not release postvocalic plosives especially /t/ and /d/ most of the time. This is why English postvocalic plosives are harder to detect automatically than prevocalic plosives.

Perhaps mothers or schoolteachers are made of sterner stuff in France because French postvocalic plosives are *always* released. Combined with the fact that French prevocalic voiceless plosives, unlike their English equivalents, are almost never aspirated, the result is that there is essentially no difference between corresponding prevocalic and postvocalic plosives in French.

Korean, as we have said, is at the other extreme: postvocalic plosives are *never* released, making them consistently different from Korean prevocalic plosives.

¹ In Korean, it is widely accepted that /y/ and /w/ are considered to be part of a diphthong, in which case the maximum number of prevocalic consonants allowed in Korean is one. However, one could also interpret these glides as consonants, in which case they are the only second prevocalic consonants that are allowed.

There is a somewhat similar pattern with the consonants /l/ and /r/. In English, there is usually a difference between the way a prevocalic and postvocalic /l/ is produced, while in French there is little if any difference. For /r/ in English, the difference between the prevocalic and postvocalic form is much more marked than for /l/. A prevocalic /r/ in American English is a definite consonant, classified as an *approximant*, while a postvocalic /r/, at least away from the northeastern seaboard of the US, is manifested as a modification of the preceding vowel. French maintains its pattern of close similarity between the prevocalic and postvocalic forms, both being a voiced fricative or approximant for most speakers. Korean is again at the other extreme, since /l/ and /r/ are effectively the postvocalic and prevocalic versions respectively of the *same phoneme*.

None of these differences renders the Novauris approach to list access inapplicable in Korean or French. Indeed, the simpler syllable structures in these languages make the overall recognition task rather easier. The message of this section, however, is that it is important to take account of the special phonetic properties of each language being addressed.

4. The Pronouncing Dictionary

Another important factor in accurate speech recognition is the quality of representation of pronunciations in the dictionary it uses. Novauris has approached this by analyzing a large corpus of training material using an automatic phonetic decoder. The pronunciations of common words can then be found by examining the corresponding outputs from the decoder. Often, this leads to multiple alternative pronunciations.

To minimize the explosion of alternative pronunciations that can occur especially in long words when multiple parts can vary, Novauris employs a compact *multiphone* notation in its dictionary. Thus instead of needing two separate entries for the two standard acceptable pronunciations of “either”, we can write the first vowel with the multiphone iy1-ay1, showing that the vowels written iy and ay are equally acceptable and that both are stressed. Moreover, a simple extension of the notation accounts for optional phonemes. Thus, the optional palatal in the word “new” can be expressed as: n -y uw1.

To cover the Novauris applications in areas including US names and addresses, popular music titles and artists, and TV programs, we currently require a dictionary with over 322,000 individual words. There is, of course, never enough material to get statistically reliable samples of the pronunciations of all these words. Instead, we work by generalization, taking common sub-parts of words (e.g. “borough” in place names) and generalizing across multiple words containing each sub-part. Web searches can also help to provide clues to how, for example, people pronounce their own names.

Finally, there are phonetic phenomena that occur across word boundaries, especially for common word pairs (“got to”, “don’t know”, *etc*). We believe that it is necessary to model such phenomena. Indeed, in addition to the individual words the Novauris dictionary lists one or more pronunciations for over 93,000 common word pairs and multi-word sequences.

5. Performance

It is natural for practical speech technologists to ask what if anything is the payoff in terms of recognition accuracy for the application of the phonetic analyses described in this paper. Since our earliest systems were sensitive to syllable structure, it is hard to say quantitatively how much we benefit from it. However, a recent extension in application of syllable-related properties, combined with the improvements to the pronouncing dictionary achieved by the pronunciation analyses described in the previous section reduced our error rate on a task of recognizing items from a list of 2.2 million US street-city-state entries (*e.g.* “Main Street, Cambridge, Massachusetts”) by a factor of more than two. The error rate is currently measured at 3.6%, falling to 1.5% when a second guess is allowed. Given the much lower redundancy of this task compared with the full name and address task that we first worked on, we feel that the low error rate we now achieve justifies our emphasis on phonetic, and especially syllabic, properties.

References

1. Kurt E. Dusterhoff, “Reducing Structure in Spoken Dialog Applications,” *Proc. SpeechTEK*, Fall Meeting, New York, 2003.
2. Melvyn J. Hunt, “Speech Recognition, Syllabification and Statistical Phonetics,” *Proc. International Conference on Speech and Language Processing, Interspeech-04*, Jeju Island, Korea, 2004.
3. Ho-Min Sohn, *The Korean Language*, Cambridge Language Surveys, Cambridge University Press, 1999, ISBN 0 521 36123 0 (hardback), 0 521 36943 6 (paperback).