Whole-Word Phonetic Distances and the PGPfone Alphabet

Patrick Juola †‡
†Department of Experimental Psychology
Oxford University UNITED KINGDOM
patrick.juola@psy.ox.ac.uk

Philip Zimmermann ‡
‡Boulder Software Engineering
Boulder, Colorado USA
prz@acm.org

ABSTRACT

Like many cryptosystems, PGPfone[13] requires a method of reliably exchanging binary data over noisy phone lines. This paper describes a method of encoding binary data into a "radio alphabet," using a feature-based distance metric to measure phonetic confusibility, then using this metric in a GA to select appropriate words from a larger list of candidate words. This work indicates several larger issues that should be addressed in any (human) language engineering project.

1. Motivations

PGPfone[13], developed by Boulder Software Engineering, provides high-quality secure voice communications over ordinary phone lines. Implicit in this project, as in any cryptographic project, are several situations where it is necessary to exchange various strings of near-random binary numbers in a secure and reliable fashion. Unfortunately, reading these strings in binary (or even hexadecimal) is tedious and error-prone.

PGPfone's designed solution to this problem is to develop a word list, styled after the traditional military or pilot's alphabet (alpha, bravo, charlie, ...), with each word representing some fixed number of bits. In addition to providing compression, this also provides some error prevention, error detection, and human factors advantages if done properly. An ideal word list would consist of short, easily recognizable and easily pronounceable words with easily distinguished prefixes and a minimum of phonetic confusibility or bad associations.

We chose to approach this task as a selection problem—from a much larger list, select words with appropriately chosen characteristics. For this project, we used the Moby Pronunciator database of nearly 200,000 word/pronunciation pairs. In some characteristics, such as "short", selection is trivial. In others, such as "no bad associations", this is nearly impossible to perform automatically and it was recognized that this would need to be done by hand. The main technical difficulty that we considered to be solvable by computer occurred in the representation of "phonetic confusibility."

2. Linguistic distances

Our approach to the problem of phonetic confusibility is a variant of the work of [9]. In particular, words are ordered strings of phonemes instead of acoustic signals, phonemes in turn contain features [such as those enumerated in [6]], and individual phonemes can be meaningfully compared by comparing their features. It is further assumed that the phonetic distance between two words can be approximated by some function of the differences between the phonemes that comprise the words.

It should be noted that this is only one of many possible approaches. In genealogy, the Soundex algorithm clusters similar-sounding names by deleting similar, repeated, or unimportant letters. Later, [7] measured vowel similarity in speech based on formant frequency. Other approaches have been proposed using "autosegmental phonology" [5] to compress word representations into feature change sets, at the expense of synchronization data. [9] and more recently [1] use a more introspective/scientific approach than simply calculating mathematical distances, instead directly examining people's perceived distances, which may or may not exactly map onto a feature-based metric—but this approach requires either extensive lab-work to validate, or a willingness to rely on pure introspection without validation.

There are several advantages to the chosen approach. First, words can be represented symbolically rather than as sound signals. Second, phonetic features (in gross) are largely uncontroversial. Third, they are usually speaker-independent, and, fourth, can be easily generalized to new words. On the other hand, there are severe representational difficulties at a number of levels. Phoneticians usually use feature sets designed to represent differences important to the production of a sound. The amount of detail, and hence importance, thus changes with the degree of variability in a feature. Sounds with voiceless stop consonants can be produced at many different locations ranging from the lips to the very back of the throat. Voicing, on the other hand, is either present or absent; no known language makes a three-way distinction among consonants. However, [9] indicate that voicing is one of the most salient and robust features of English con-

Feature name	Sample	#bits
Place of articulation	/d/ vs /g/	7
Manner of articulation	/l/ vs /t/	6
Height of articulation	/i/ vs /ϵ/	5
Voicing	/z/ vs /s/	4
Syllabic	vowels vs. cons.	1
Nasal	/n/ vs /d/	1
Lateral	/l/ vs /r/	1
Roundedness	(various)	1

Table 1: Phoneme coding for PGPfone alphabet

sonants; generalizing this yields the unfortunate result that phoneme pairs may differ in several unimportant features and yet sound closer than another pair that differ only in one extremely salient aspect.

Furthermore, the relative salience of features varies wildly depending upon the sort of noise in which the signal is embedded. Finally, although [9] provide exact data that could be used to balance some features, they don't provide enough. [6] proposes a more extensive list of features that allow for all sounds of English, consonants and vowels alike, to be presented and distinguished on a uniform scale but provides no data on salience. With appropriate judgements, the various features can be approximately balanced to the data. Using this method, the perceptual difference between two comparable phonemes can be measured as the number of bits that differ in the two representations. The final representation is attached as table 1.

Even granting the viability and success of a phoneme-by-phoneme perceptual distance metric, there are difficulties in its extension to full-word distances, and here theory provides less support than might be wished. For example, if each phoneme were weighted equally and could be directly compared with a single other phoneme, the difference between two words is simply the sum of the phoneme differences. However, some phonemes are clearly more salient than others. On a gross level, the stressed syllables of a word pair are intuitively of much greater salience than the unstressed ones. Furthermore, psycholinguistic results such as [4, 11] suggest that onsets are more salient than codas. The approach taken in PGPfone was a simple one; the preceding consonant cluster and vowel(s) of the stressed syllable were given twice normal weight, as was the (word-)onset phoneme.

A similar problem arises with non-aligned or non-existent sounds. For example, should the word /bEst/ be treated as most similar to /bEts/, /bEt/, or /bEs/? [4] present a few primitive metrics to address this question, based on primarily on a notion of sequences of identical vs. nonidentical phonemes. A more sophisticated approach could use the notion of "edit-distance" as typified by [10], but only with an accurate measure of the perceived difference between a sound and its absence, or in other words, a featural representation for silence.

This problem can be reduced by the use of templates. For instance, if all the words in a study are of the form CVC, then there need be no representation of silence as all phonemes are aligned directly. For independent reasons (discussed below), the PGPfone list demands words with small consonant clusters and of a particular syllabic structure. The list to be selected from was filtered before the selection process began to eliminate unsuitable words with long consonant clusters. By increasing the strength of the prefiltering, one can restrict attention to words where comparisons are meaningful, or phrased another way, one can greatly limit the damage done by a bad representation of silence. Similarly, by careful use of duplicate sounds, some of the silence/sound comparisons can be avoided. Vowel sounds can be lengthened or shortened almost at will, thus, vowel blends (such as /Oi/) are compared with "pure" vowels such as /i/ by the simple technique of presenting the pure vowel twice (/ii/) and comparing—and thus /Oi/ is accurately represented as midway between /O/ (/OO/) and /i/ (/ii/).

One final concern for the PGPfone distance metric touches on the incorporation of additional, non-linguistic features. For example, it would be nice if the final words had distinguishable orthographic prefixes, to make it easier for keyboard entry and similar (non-linguistic) processing tasks. Obviously, paying attention to such things will, in theory, weaken the linguistic quality of the final solution but result in a better system overall.

3. Engineering Aspects

The ultimate test of any representation is the quality of solutions it permits. A good solution for the PGPfone list involves more qualities than simply an accurate distance representation, as detailed in this section.

The most obvious engineering aspect is the length of the word list to be developed. A small list (for example, sixteen words) would provide no compression over reading hexadecimal numbers, but could still provide some degree of error-proofing by removing potentially confusing tuples such as five/nine, B/C/D/E, and so forth. A list of 64 words would allow about 30% compression, but be harder for a human to memorize. For situations where humans are required to generate keys from memory, this would be unreasonable. However, in PGPfone, all keys are generated and stored by the computers, and the only job for a human is to read a series of words presented by the computer; thus, there is no need for a human to memorize the list. Using larger lists would allow better compression, but require more (computer) memory to store the word table. For example, two lists of 256 words can be stored in only 5 kilobytes of memory. A larger list (two bytes per word) would require nearly 650 kilobytes of memory, as well as a word vocabulary larger than most speakers' vocabulary.

Humans, when reading sequences, tend to make different errors from simple bit-flips (misreadings), so error detection

and recovery is not simply correcting bits. [12] suggested a clever way to allow human-like errors to be easily detected. By building *two* lists instead of one, and alternating the lists from which the words in the sequence come, one can easily spot common human errors (omission, repetition, and transposition) by noticing that two successive words come from the same list. This assumes, of course, that the (listening) human can tell from which list a word has been drawn.

Similarly, the lists should consist of words that are easily pronounceable and easily readable. Thus, words with multiple spellings or multiple pronunciations, words for which we had evidence of significant dialect variations (e.g. tomato), or any hard-to-pronounce words, defined as words incorporating any non-English sounds or lengthy consonant or vowel clusters, were also eliminated.

The most difficult aspect of the list to control was unfortunately one of the more important; the final lists should contain words with appropriate associations. One of the goals was to develop a word list that would inspire a certain amount of confidence in the security of the overall product. Although it proved difficult enough to automatically banish repugnant words there seems no automated procedure for detecting all and only "cool" words. We used what ad hoc principles we could identify. The standard pilots' alphabet, for instance, contains familiar but uncommon words; only a few words do not appear in the Brown corpus at all, while no word listed has a frequency of 85 occurrences or more. So words that were too unfamiliar or too common were eliminated. In general, noninflected words seem stronger than their inflected variants. In the end, we were forced to rely on human judgement, generating a list, blue-pencilling or modifying words that we found inappropriate, then using the survivors as the base for another list.

Once the selection and measuring criteria are available, the actual selection of the list is, technologically speaking, near-trivial. Because, as discussed above, only the computer will ever need the full list, we opted to use lists of 256 words, allowing each word to represent a byte. The lists for PGPfone are obviously different in that one consists only of two syllable words, and the other, three. We used a simple genetic algorithm [2] to evolve a near-optimum (sub)set of the candidates such that the smallest distance between any pair was maximized. Specifically, the GA generated a population of random 256-word subsets of the candidate list. Subsets were permitted to "breed" by trading some of their members, and the daughter subsets were evaluated (using the distance metric described above) to determine the closest pairwise distance. Successful children were allowed to be fruitful and multiply, while the losers in the genetic sweepstakes were simply dropped from the population. After several hundred generations, the top candidate was then edited as described above, and the surviving words were used as a fixed and unchanging part of the entire population for the next run of the GA selection program. Because our automated selection procedure relied on two separate lists, using two separate and incomparable templates, we also performed a form of cross-checking between lists to assure significant differences between the two lists.

111	glucose	hesitate	116	guidance	impartial
112	goggles	hideaway	117	hamlet	impetus
113	goldfish	holiness	118	highchair	inception
114	granny	hurricane	119	hockey	indigo
115	gremlin	hydraulic	120	hotdog	inertia

Table 2: Sample words from the PGPfone list

After several runs, when a final, accepted list had been agreed upon, the words in each list were alphabetized without regard to case and used to represent byte values from 0 to 255. Some sample words from the middle of the lists are here attached as table 2.

4. Implications and Conclusions

The final alphabet as distributed in PGPfone appears to work well enough for the purpose for which it was designed; our feedback has generally been positive, and suggested improvements tend to be matters of opinion on single words rather than changes to the underlying structure or model. Figure 1 shows how this alphabet can convert a relatively unmnemonic key fingerprint into something more memorable (and easier to confirm over a telephone).

This work does strongly suggest the need for further work on the development of word-scale phonetic confusibility models. The alphabet itself might have been made much stronger if we had been able to take several dozen subjects into a phonetics laboratory and test the weightings conjectured above. Fundamental data on the salience of various word-level characteristics is available only in a very sketchy manner (and likely to vary significantly with language anyway.)

Clearly, a full evaluation of this work requires some empirical checking, which at this point has not yet been done. Although informal tests show that the words are understood, the degree of confusibility has not been rigorously tested. There are many open questions. How confusible *are* the words? Does the actual transmission channel fit the assumptions? Do the pronunciation assumptions fail when the reader is not a native English speaker?

Although this problem may seem artificial in many regards, it lends itself well to treatment as a touchstone problem for many speech/language generation problems. The difficulty we encountered with the representation of consonant clusters mimics the difficulties other researchers such as [3, 8] have had with the learning and representation of sound patterns in language acquisition tasks. Particularly in situations such as neural networks or supervised learning, where a distance measure is used to direct the system to its new state, an accurate distance measure is more a necessity than a convenience. An accurate statistical analysis of the effectiveness

```
-----BEGIN PGP PUBLIC KEY BLOCK-----
Version: 2.6.2i

mQCNAzDIPcEAAAEEALvWEowkZJ8sLUnOcMkCykWpjKirlwEv3LAC6c6ciU63bhzn
yVcH22KKZQj6n+A2sIn+qLdKiKlLNOdOBh7wIwlJrlYb/g6zMyw6TpWPPRopzkis
7U2eofSKZ4L19RSVw8+QejFvHeMx89+QdTzUNXTAAthkJZporyC+v3X+p5ZhAAUR
tCpQYXRyaWNrIEp1b2xhIDxwYXRyaWNrLmp1b2xhQHBzeS5veC5hYy51az6JAJUD
BRAw5a+WZXmEuMepZt0BAfSdA/0VQcMu1oV8N1Tx4MTI8gk/FN7BYH5PHFpFOQrQ
Ahr4NZKN395q7LvMPb6jbsuLAI2eamg6ujQZU3X5iiXMS58dm7F7ATz0PRVh9768
dl62STyMNVMBbYc2Wqruk7jDHIw2HaU+8CMSWJE66FbKO8y7TnAy1TTIXOvHR6OL
EOWLbw==
=uB9I
-----END PGP PUBLIC KEY BLOCK-----

Key fingerprint =  5C 39 76 D5 DD E2 9E C2  56 2C A2 91 C7 91 65 F9
Encoded fingerprint =
escape crossover hotdog speculate swelter torpedo puppy reproduce egghead
combustion quota molecule spaniel molecule fracture Waterloo
```

Figure 1: Patrick's PGP key, with encoded fingerprint

and salience of various feature-based representations may shed light to bridge the sound/phoneme gap—as well as help with the (word) segmentation problem and provide fundamental evidence about the psychological reality of phonemes and phonetic features. From an engineering perspective, an accurate way of measuring perceptual distance could help in any situation where language must be engineered or corrected, for example in sublanguage design, speech-text conversions, database retrieval, or more prosaically to help with the creation of distinctive product names.

This work illustrates several basic principles that a reasonable metric should follow:

- First, that standard feature sets do not accurately reflect the perceived salience of various features.
- Second, that feature differences are a significant but not all-encompassing part of the perceived differences among words; superphonemic attributes such as stress and onset must also be taken into account.
- Third, that templates are best used to control the sorts of comparisons and measurements taken, but that using them will greatly restrict the overall validity of the measurements.

There are almost certainly other principles that could be found and added to this list. It is hoped that future work, whether in the context of PGPfone 2.0 or other unrelated projects, will be able to extend this list of principles to a full theory of isolated word perception.

5. REFERENCES

1. Alan Bell. Personal communication, 1995.
2. Albert Donally Bethke. *Genetic Algorithms as Function Optimizers*. PhD thesis, University of Michigan, 1981.
3. Garrison W. Cottrell and Kim Plunkett. Acquiring the mapping from meaning to sounds. *Connection Science*, 6(4):379–412, 1994.
4. Bruce L. Derwing and Terrance M. Nearey. Experimental phonology at the University of Alberta. In John J. Ohala and Jeri J. Jaeger, editors, *Experimental Phonology*. Academic Press, Inc., Orlando, Florida, 1986.
5. John A. Goldsmith. *Autosegmental Phonology*. PhD thesis, Massachusetts Institute of Technology, 1976.
6. Peter Ladefoged. *A Course in Phonetics*. Harcourt Brace Jovanovitch, London, 3rd edition, 1993.
7. Bjorn Lindblom. Phonetic universals in vowel systems. In John J. Ohala and Jeri J. Jaeger, editors, *Experimental Phonology*. Academic Press, Inc., Orlando, Florida, 1986.
8. Brian MacWhinney. Lexical connectionism. In Peter Broeder, editor, *Cognitive Approaches to Language Learning*. MIT Press, Cambridge, Mass., 1995.
9. George A. Miller and Patricia E. Nicely. An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, 27(2):338–52, 1955.
10. Eugene W. Myers. An O(ND) difference algorithm and its variations. *Algorithmica*, 1:251–56, 1986.
11. Dan Isaac Slobin. Crosslinguistic evidence for the language-making capacity. In Dan Isaac Slobin, editor, *The Cross-Linguistic Study of Language Acquisition*, pages 1157–1256. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1985.
12. Zhahai Stewart. Personal communication, 1991.
13. Philip Zimmermann. *PGPfone Owner's Manual*. Boulder Software Engineering, Boulder, Colo., 1995.