

## Phonetic Similarity Costing

The STANDUP joke generator compares two strings of phonetic symbols (the representation of word forms) by applying a Levenshtein minimum edit-cost algorithm, normalised for length. However, within the standard edit-cost algorithm, the cost of substituting one symbol for another is not constant, but depends on how different the two symbols are – very similar symbols incur a low cost, dissimilar symbols cost more. To have a plausible estimate of the costs, we used the feature decomposition of the IPA symbols in [1] as a starting point, and formalised it as outlined below.

The framework involves various *attributes*, such as **Voicing**, **Height**, etc., each of which has a set of allowable values; for example, **Voicing** can be **V** (‘voiced’) or **U** (‘unvoiced’). The attributes are in three *Levels*, where Level 1 has only one attribute, **VC** (distinguishing vowels from consonants), Level 2 has 6 attributes (3 for classifying vowels, 3 for consonants) and Level 3 has various fine-grained features, as described below. Each attribute has an associated *cost*, a value in [0,1], which was our intuitive guess at how much difference this attribute made to the similarity of two phonetic symbols. Each Unisyn symbol was associated with a set of *features*, where a feature is an *attribute-value* pair. Two symbols are costed at the highest level at which they do not have identical feature values. That is, if Unisyn symbols  $S_1$  and  $S_2$  have identical Level 1 features, then their Level 2 features are considered, and so on. At the level used for costing (i.e. the highest at which a difference exists) the cost is computed as the total of the costs associated with those attributes for which the symbols have different values.

**Level 1:** One attribute, distinguishing vowels and consonants (possible values in braces):

**VC:** {V, C}

**Level 2:**

For [VC:V] at Level 1:

**Frontness:** {F,C,B}

**Height:** {0,MO,MC,MCC,C,OC}

**Rounding:** {R, U, N}

For [VC:C] at Level 1:

**Voicing:** {V,U}

**Place:** {BL, LD, LV, D, A, PA, P, V, G}

**Manner:** {SP,A,N,F,A,LA,FL}

**Level 3:** Each particular combination of Level 2 features defines a very narrow class of symbols; for example {**Frontness:F**, **Height:0**, **Rounding: U**} is the class of front open unrounded vowels. In Unisyn, such classes are not always singletons; for example, the three symbols **@**, **@r**, **@@r** (variants on schwa)

share the marking {**Frontness:MC**, **Height:C**, **Rounding:U**}. To distinguish members of these narrow classes, there is a unique Level 3 attribute for each such class (i.e. for each combination of Level 2 features); e.g. **F-0-U** for front open unrounded vowels. This attribute has one possible symbolic value for each Unisyn symbol in that narrow class, so that each Level 3 attribute-value pair (e.g. **MC-C-U:@r**) corresponds to a unique Unisyn symbol.

Hence a symbol has exactly one Level 1 feature, three Level 2 features, and one Level 3 feature. That is, most of the real distinctions are at Level 2.

**Costs:** Costs are allocated following certain postulates:

- Not only are vowels and consonants phonetically quite dissimilar, vowel-consonant substitutions tend to disrupt syllable structure, whereas vowel-vowel or consonant-consonant substitutions tend to preserve structure. So VC costs 1.0 (maximum dissimilarity).
- Substituting a consonant for a consonant will cost slightly more than a vowel for a vowel, given comparable levels of dissimilarity. Intuitively, consonants indicate the structure of the word. All Level 2 consonant attributes are costed at 0.28, vowel attributes at 0.15.
- Symbols which differ only at Level 3 should have extremely small costs for substitution, as they are almost identical. Level 3 attributes are costed at 0.15.

Tables showing the allocation of Level 2 features to Unisyn symbols are given below.

## References

- [1] Peter Ladefoged and Morris Halle. Some major features of the international phonetic alphabet. *Language*, 64(3):577–582, 1988.

Symbol	H	F	R
@	MC	C	U
@@r	MC	C	U
@r	MC	C	U
a	O	F	U
ae	O	F	N
aer	O	F	N
ai	O	F	N
ar	O	F	U
e	MO	F	U
ei	MC	F	U
eir	MC	F	U
er	MO	F	U
i	MCC	F	U
ii	C	F	U
ii;	C	F	U
ir	C	F	U
ir;	C	F	U
oi	MO	B	R
oir	MO	B	R
oo	M	B	R
or	M	B	R
ou	MC	B	R
our	MC	B	R
ow	OC	B	U
owr	OC	B	U
uh	MO	B	U
ur	C	C	R
ur;	C	C	R
uu	C	C	R
uu;	C	C	R

Table 1: Vowel Level 2 Features

Symbol	V	P	M
?	U	G	SP
b	V	BL	SP
ch	U	PA	A
d	V	A	SP
dh	V	D	F
f	U	LD	F
g	V	V	SP
h	U	G	F
hw	U	LV	A
jh	V	PA	A
k	U	V	SP
l	U	L	F
l!	V	A	FL
m	V	BL	N
m!	V	LD	N
n	V	A	N
n!	V	P	N
ng	V	V	N
p	U	BL	SP
r	V	PA	F
s	U	A	F
sh	U	PA	F
t	U	A	SP
t^	V	PA	FL
th	U	D	F
v	V	LD	F
w	V	LV	A
x	V	V	F
y	V	P	A
z	V	A	F
zh	V	PA	F

Table 2: Consonant Level 2 Features