Some measures of phonetic similarity for use in legal trademark disputes

Sandra Ferrari Disner[1] & Vincent J. van Heuven[2, 3, 4]

[1] University of Southern California, Los Angeles, USA
[2] Leiden University Centre for Linguistics, Leiden, The Netherlands
[3] University of Pannonia, Veszprém, Hungary
[4] Fryske Akademy, Leeuwarden, The Netherlands
s.disner@usc.edu; v.j.j.p.van.heuven@hum.leidenuniv.nl

ABSTRACT

A trademark or brand name may be legally disputed if it sounds too similar to another trademark. US courts would benefit from simple and understandable measures of sound-alikeness that judges and lay jurors would be able to understand. We compared the usefulness of transitional probabilities (bigrams and trigrams in phonetic transcriptions) and various string edit measures (Levenshtein distance, length-normalised or not, feature-weighted or not) in two studies. In the first, we computed similarity and distance measures for all US trademark litigation cases we could find that were settled predominantly on the basis of sound similarity. We aimed to find optimal cut-off values to separate the names that were deemed too similar from those deemed sufficiently distinct and thus allowed to compete in the market. Our second study examined 120 pairs of generic drug names used in the USA that were deemed confusable by pharmaceutical experts. Again, length-normalized and feature-weighted Levenshtein distance proved the best predictor of confusability.

Keywords: trademark sound-alikes, shared n-grams, Levenshtein string edit distance, phonetic feature weighting, string length normalization.

1. INTRODUCTION

A trademark is a type of intellectual property consisting of a recognizable sign (e.g., logo), design (characteristic shape, size, color), and/or linguistic expression, which identifies and distinguishes products or services of a manufacturer or provider from those of potential competitors. It happens with some regularity that a new brand, e.g., *Americrest* (a mortgage lender) on the market semi-copies the trademark of an earlier, well-established product of the same type, e.g., *Ameriquest*, in an attempt to derive unearned benefits from the reputation of the 'senior' mark ('reputation parasitism'). Although such stealing of intellectual property is illegal, no *a priori* checks are mandatory for an entrepreneur who wants to penetrate the market with a new trademark. Any initiative for starting litigation, and the burden of proof, is on the owner of the senior mark, who must show to the satisfaction of a court of law that the junior mark bears a confusing similarity to the senior mark, and may be held accountable for (potential) loss of product turnover. In such cases, forensic phoneticians may be called upon as experts to argue that the sound shapes in a trademark dispute are (or are not) similar enough to cause confusion among consumers [1, 2, 3].

Generally, in such cases, the court does not allow litigants to run (field) experiments with human participants to determine the degree of auditory confusability between trademarks. The reason for this is that the confusability of competing brand names is strongly affected by the articulation and voice characteristics of the speaker, the auditory acuity and motivation of the listener, and on the noisiness of the communication channel. Rather, the court wants theoretically grounded, generally applicable reasoning to establish the degree of confusability that can be explained to (and understood by) a lay jury. In current judicial practice, the court's decision whether two contested trademarks are sufficiently distinct or sound so similar that auditory confusion may arise, are made on intuitive grounds, and are typically not based on systematic phonetic reasoning. But here we explore the possibility of more objective methods to determine auditory confusability in legal trademark disputes. We will do this in two background studies. However, before we present summaries of these studies, we will first explain techniques used in the literature, and by ourselves, to quantify the degree of similarity or difference between two sound shapes, be they existing words or nonce trademarks.

2. MEASURES OF SIMILARITY

Characteristic of all the measures we will survey in this paper is that they are not computed directly on some acoustic signal recorded from a human speaker. We abstract away from a specific human speaker by converting the sound shapes to a phonetic transcription, which may be broad (one symbol per phoneme) or somewhat narrow, marking predictable allophones with dedicated transcription symbols or diacritics. In our studies on American English trademarks so far,

we have used standardized broad transcriptions only, including the primary stress, which we treat on a par with the segmental phonemes.

Measures of similarity between strings of transcription symbols are typically based on the number of symbols shared by two words or names [1, 2, 3]. To preserve information on sequential order, measures of similarity are based on the number of n-grams shared by two sound shapes, where $n > 1$ [4]. In most studies the n-grams are either bigrams (sequences of two adjacent symbols) or trigrams (sequences of three adjacent symbols). A word boundary symbol '#' is inserted at the beginning and end of the transcription of a name/word, and is counted on a par with the other symbols, to preserve information about the edge status of the first and last sound of a word. The bigram and trigram measures are generally correlated but provide partly independent information about similarity between two strings of symbols. A general concern in expressing the degree of similarity (and difference) between two sound shapes is the length of the strings.

The longer the strings are, the better the chances of the same n-gram occurring in both strings. Rather than counting the absolute number of shared n-grams, therefore, we compute the percentage of shared n-grams. Any n-gram that occurs in both string A and string B, increments the count by 2. The sum of the shared n-grams is then divided by the total number of n-grams in A and B added together. Table 1 shows how to compute the percentage of shared bigrams and trigrams (Pbi, Ptri) for the brand names *Phexxi-Imvexxy* /f 'ɛksii/-/ɪmv'ɛksii/, i.e., 67 and 63, respectively.

Table 1: Computation of shared n-grams.

	Bigrams			Trigrams		
	f 'ɛksii	ɪmv'ɛksii	*N*	f 'ɛksii	ɪmv'ɛksii	*N*
1.		#ɪ	0		#ɪm	0
2.		ɪm	0		ɪmv	0
3.	#f	mv	0	# f '	mv'	0
4.	f '	v'	0	f 'ɛ	v'ɛ	0
5.	'ɛ	'ɛ	2	'ɛk	'ɛk	2
6.	ɛk	ɛk	2	ɛks	ɛks	2
7.	ks	ks	2	ksi	ksi	2
8.	si	si	2	sii	sii	2
9.	ii	ii	2	ii#	ii#	2
10.	i#	i#	2			
	Σ shared bigrams		12	Σ shared trigrams		10
	Σ bigrams		18	Σ trigrams		16
	Pbi = (12/18) × 100		67	Ptri = (10/16) × 100		63

In our broad transcription, long/tense vowels are represented as geminates, including the vowels in *bad* /ææ/ and *hot* /ɑɑ/. Diphthongs in *find* /aɪ/, *found* /aʊ/ and *coin* /ɔɪ/, as well as the diphthongized tense vowels in *fame* /eɪ/ and *foam* /ou/ are transcribed and computationally treated as two vowels in sequence.

3. MEASURES OF DIFFERENCE

Differences between strings of symbols are based on string edit counts. The Levenshtein distance (LD) measure was first proposed by the eponymous Levenshtein [5], and later adopted in the field of computational dialectology as a convenient and valid measure of the distance between related varieties (dialects, accents) of a language [6, 7]. LD counts the number of string edit operations needed to convert a string of symbols A to its counterpart B. Possible string operations are: insertion, deletion and substitution of a symbol [8]. LD software is available on the internet [9, 10, 11]. Before strings can be compared, they have to be optimally aligned through dynamic programming. Vowel symbols are aligned with vowel symbols, consonant symbols with consonant symbols. Semivowels /j, w/ can be aligned with vowels as well as consonants. Similarly, the neutral vowel schwa can be aligned with any other vowel or with /r/. The LD algorithm minimizes the cost of the conversion by finding the optimal alignment of symbols and the least number of edit operations. In the binary (or 'plain') application of the algorithm, each edit incurs a penalty of 1 point. Some implementations of the algorithm allot .5 penalty to either an insertion or a deletion (also called 'indel') and 1 penalty point to a substitution. Operation on a diacritic in a narrow transcription incurs a cost of .5. Optional length normalization is achieved by expressing the percentage of the raw LD relative to the maximal cost that could be incurred by the two strings under comparison, where the max cost of a substitution equals 1, and of an indel .5. Table 2 exemplifies the computation of LD for the string pair /f 'ɛksii/-/ɪmv'ɛksii/, assuming a broad transcription as input.

Table 2: Computation of raw and length-normalized Levenshtein Distance (LD).

			Binary (plain)		F-weighted	
	f 'ɛksii	ɪmv'ɛksii	cost	max	cost	max
1.		ɪ	.5	.5	.39	.5
2.		m	.5	.5	.50	.5
3.	f	v	1.0	1.0	.09	1.0
4.	'	'	0	1.0	0	1.0
5.	ɛ	ɛ	0	1.0	0	1.0
6.	k	k	0	1.0	0	1.0
7.	s	s	0	1.0	0	1.0
8.	i	i	0	1.0	0	1.0
9.	i	i	0	1.0	0	1.0
	Raw LD		2.0	8.0	.98	8.0
	Norm. LD (%)			25		12

It has been objected that the binary LD is too crude a measure, as the substitution of /f/ for /v/ raises the cost no higher than the substitution of, e.g., /t/ for /w/. Though no improvement could be demonstrated [6], judge and jury will be more easily convinced of the validity of the LD measure when the costs are weighted for the number of distinctive features that differ between the two sounds involved in the substitution. The LD algorithm we use optionally determines the difference between two vowels, or two consonants, according to the number of phonetic features that have to be changed to convert one segment to the other. It uses the feature system proposed by [12], and adapted for use in the LD computation by [13]. Changing /f/ to /v/ affects one consonant feature, [voice], which incurs a small cost of only .09 (see Table 2 under F-weighted, i.e., feature weighted). For reasons of space, we do not include the weighted cost matrices for vowels and consonants but will make these available in the supplementary materials. We also refer to the help files included in the LED-A software [11, 14].

4. STUDY 1: PREDICTING COURT DECISIONS

In our first study we examined the first 200 trademark cases in the USA listed in the Lexis-Nexis legal database.[1] These are the most frequently cited cases involving likelihood of confusion in the past 25 years. We kept only those that involved a dominant role of auditory confusion of the product names in the court's decision. In a second round, we searched the entire database with the conjoint terms: 'lapp test' & 'sound' & 'likelihood of confusion', which narrowed the cases down to those in which the court utilized a well-known test for the existence of likelihood of confusion based on the degree of sound similarity. The search yielded a set of 51 pairs of contested trademarks for which a court decision was known. Thirty-four contested pairs were judged too similar to compete on the market, while the remaining 17 pairs were judged sufficiently different to avoid consumer confusion.

We computed four similarity/distance measures between each of the trademark pairs as explained above. Table 3 shows the (Pearson) correlation matrix for the four measures.

Table 3: Correlation matrix for four similarity/distance measures ($p < .001$ in all cases).

	Pbi	Ptri	pLD
Shared bigrams (%)			
Shared trigrams (%)	.942		
Plain Levenshtein Distance	−.900	−.908	
Feature-weighted wLD	−.871	−.858	.920

Given the high intercorrelations, we deleted one similarity and one distance variable, i.e., those that were least successful in distinguishing pairs that had been judged too similar from those that had been judged sufficiently different. The two remaining variables, Pbi and wLD, were then entered as predictors in a Linear Discriminant Analysis (LDA [15]) distinguishing the two types of pairs. We ran the LDA three times, i.e., once with Pbi as a single predictor, once with wLD as a single predictor, and a third time with both predictors combined. Results are shown in Table 4.

Table 4: Results of LDA. *N* correct classifications in green cells.

Predictor(s)	Court's decision (down)	Predicted		Total
		TS	SD	
Pbi	Too similar (TS)	24	10	34
	Sufficiently different (SD)	3	14	17
	76.5% correct, TS ≥ 61.0% Pbi			
wLD	Too similar (TS)	28	6	34
	Sufficiently different (SD)	4	13	17
	80.4% correct, TS ≤ 14.5% wLD			
Pbi+wLD	Too similar (TS)	29	5	34
	Sufficiently different (SD)	4	13	17
	82.4% correctly classified boundary: 1.251×z(wLD) + .322×z(Pbi) = 0			

In Figure 1, we illustrate the contribution of wLD and Pbi to the correct separation between 'too similar' vs 'sufficiently different'.

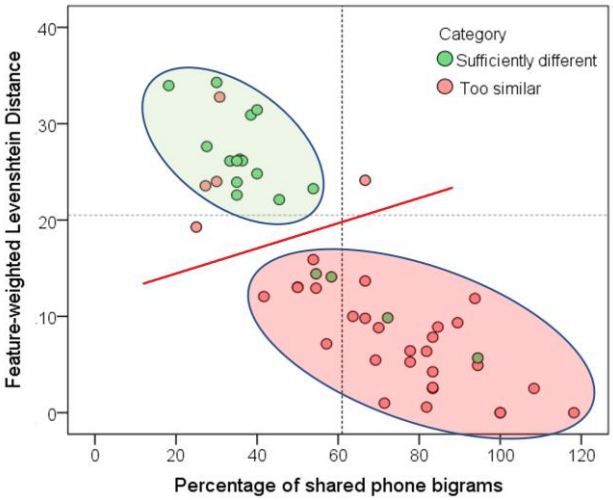

Figure 1: Separation of contested pairs of trademarks by wLD and Pbi together. Ellipses and boundary drawn by hand.

Some of the error decisions (red dots in green area, or green dots in red area) seem to be the result of the court's decision having been based on non-phonetic grounds after all (see [16: slide 30] for examples and discussion). Generally, the critical values for Pbi and

wLD work very well to distinguish 'too similar' and 'sufficiently different'. We ran an external validation on these parameters by computing the distribution of Pbi, Ptri, pLD and wLD for all possible pairs of English mono-morphemic words taken from the 3000 most frequent lexemes according to the British National Corpus but using the American English phonemic transcriptions in the CMU digital pronouncing dictionary [17]. In the 4,305,656 unique word pairs, a Pbi $\leq$ 61% is seen in fewer than .03% of the pairs. A wLD < 14.5% is found in 1.2% of the pairs. The conclusion follows that trademarks have to be exceptionally similar before they will be banned by a US court decision.

5. STUDY 2: PREDICTING CONFUSABLE DRUG NAMES

In our second study we did not attempt to predict (or rather postdict) the court's decision on competing trademarks; rather, we addressed the issue of consumer confusion more directly, by looking at confusable (generic) pharmaceutical drug names.[2] About 25% of all medication errors can be ascribed to some confusion of drug names [18], which may cause the issuing of a wrong prescription because of incorrect recognition of a drug name (either by eye or by ear). We were given a list of 1,250 pairs of drug names that can be obtained in the USA, which were deemed confusingly similar by a panel of pharmacologists.[3] We intended to use these pairs to see how confusable pairs could be automatically identified, using the parameters we found in the previous study.

The first problem we had to solve was to establish the official pronunciation of the drug names concerned. Here we should distinguish between branded drug names and generic drug names. The pronunciation of branded drug names is up to the manufacturer, and cannot easily be determined. The pronunciation of generic drug names, however, is laid down by the United States Adopted Name (USAN) Council, and can be obtained online from the USP dictionary of USAN and International Drug Names.[4] The pronunciation is specified in a laymen's phonetic transcription, which includes both primary and secondary stresses (sometimes debatable) and uses orthographic letter combinations that can only be pronounced in the way the USAN council wants the names to sound. We selected only the generic drug names from the list of 1,250 pairs, and within this subset we eliminated all compound names, i.e., drug names with elements separated by spaces, hyphens, numbers or other non-letter characters. Next, we checked in the USP dictionary if a pronunciation was listed (in which case we were certain that the name was indeed generic), copied the transcription and converted it (automatically) to the IPA transcription required by the LD software. This selection yielded 120 LASA (Look-Alike, Sound-Alike) pairs of generic drug names. The same name may occur in multiple pairs, e.g., as in *Oxazepam-Quazipam* and *Oxazepam-Oxaprozin*.

Apart from the test set of 120 LASA pairs we created a control set by generating all non-implicated pairs of these 120 drug names, i.e., a set of 2,428 non-confusable pairs. The total dataset then comprised 120 + 2,428 pairs of drug names, for which we computed 14 similarity/distance measures. Eight addressed phonetic similarity/distance:

Nbi	(raw)	Pbi	(length normalized)
Ntri	(raw)	Ptri	(length normalized)
pLD_r	(raw)	pLD_n	(length normalized)
wLD_r	(raw)	wLD_n	(length normalized)

The other measures were the same as the above but computed on the orthographic forms (printed names); however, no orthographic measures for weighted letter shapes were defined.

The 14 measures were used to classify the 2,548 pairs of drug names into those that were held to be confusable, and those that were non-confusable controls. Each measure was tested once separately, while the most successful predictors were also tested in selected combinations. Single predictors successfully separated the two categories between 83.8 and 86.4% correct for phonetic parameters, and even better for orthography-based parameters, i.e., between 82.5 and 92.3% correct. The best combination of phonetic predictors was (as before) length-normalized feature-weighted LD plus the percentage of shared bigrams Pbi (87.5% correct classification). Length-normalized predictors were always more successful than their raw (non-normalized) counterparts. Feature-weighted LD (raw as well as length normalized) proved a better predictor of confusability than the plain (binary) counterpart measures.

6. CONCLUDING REMARKS

In this paper we have tried to demonstrate that intuitive (court) decisions on similarity and confusability of consumer products can be predicted with high accuracy by relatively simple and straightforward linguistic-phonetic procedures that a layman (judge or member of a jury) would understand. Our measures and boundary values can be profitably used to signal potential consumer confusion, so that manufacturers may pre-empt legal procedures and choose names that are on the safe side of the boundaries.

More research is needed to separate look-alike from sound-alike confusability. This will require selection of critical name pairs in which sound-alikeness does not correlate with look-alikeness.

7. REFERENCES

[1] Butters, R. R. 2010. Trademark linguistics. Trademarks: Language that one owns. In: Coulthard, M., Johnson, A. (eds.), *Routledge handbook of Forensic Linguistics*. Routledge, 351–364. Doi: 10.4324/9780203855607

[2] Shuy, R. 2002. *Linguistic battles in trademark disputes*. Palgrave Macmillan. Doi: 10.1057/9780230554757

[3] Shuy, R. 2012. Using Linguistics in trademark cases. In: Solan, L., Tiersma, P. (eds.), *Oxford handbook on Language and Law*. Oxford University Press, 449–462. Doi: 10.1093/oxfordhb/9780199572120.013.0033

[4] Stephen, G. A. 1994. *String searching algorithms*. World Scientific.

[5] Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Sovjet Physics Doklady* 10, 707–710.

[6] Heeringa, W. J. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Center for Language and Cognition Groningen. https://pure.rug.nl/ws/portalfiles/portal/9800656/thesis.pdf.

[7] Kessler, B. 1995. Computational dialectology in Irish Gaelic. *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, 60–67

[8] Kruskal, J. B. 1999. An overview of sequence comparison. In: Sankoff, D., Kruskal, J. (eds.), *Time warps, string edits, and macromolecules. The theory and practice of sequence comparison*. Stanford Center for the Study of Language and Information, 1–44.

[9] Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P., Leinonen, T. 2011. Gabmap – a web application for dialectology. *Dialectologia: revista electrónica*, 65-89. https://www.clarin.nl/sites/default/files/restore/Gabmap-long-2011-jan-07-rev-mei.pdf

[10] Leinonen, T., Çöltekin, Ç., Nerbonne, J. 2016. Using Gabmap. *Lingua*, 178, 71–83. Doi: 10.1016/j.lingua.2015.02.004

[11] Heeringa, W. J. 2021. Levenshtein Edit Distance App (LED-A). Computer program. https://fryske-akademy.nl/fa-apps/led-a/#run

[12] Almeida, A., Braun A. 1986. "Richtig" und "falsch" in phonetischer Transkription; Vorschläge zum Vergleich von Transkriptionen mit Beispielen aus deutschen Dialekten. *Zeitschr. Dialektol. Ling.* 53, 158–172. https://www.jstor.org/stable/40502947

[13] Heeringa, W. J., Braun, A. 2003. The use of the Almeida-Braun system in the measurement of Dutch dialect distances. *Comp. Hum.* 37, 257–271. http://wjheeringa.nl/papers/cath02a.pdf

[14] Heeringa, W. J., van Heuven, V. J., Van de Velde, H. 2022. LED-A: a web app for measuring distances in the sound components among local dialects. Poster presented at the 17th New Methods in Dialectology Conference, Mainz. https://methodsxvii.uni-mainz.de/files/2022/06/

[15] Klecka, W. R. 1980. *Discriminant analysis*. Sage.

[16] van Heuven, V. J., Disner, S. F., Heeringa, W. J. 2021. What's in a name? On the phonetics of trademark infringement. Plenary talk presented at the 29th Annual Conference of the IAFPA, Marburg. Doi: 10.13140/RG.2.2.17529.95846

[17] The Carnegie Mellon University Pronouncing Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[18] Lambert, B. L. 1997. Predicting look-alike and sound-alike medication errors. *Am. J. Health-System Pharm.* 54, 1161–1171. Doi: 10.1093/ajhp/ 54.10.1161

[19] van Heuven, V. J., Disner, S. F. 2022. Utility of length-normalization for predicting confusion of generic drug names from Levenshtein string edit distance. Plenary talk presented at the 30th Annual Conference of the IAFPA, Prague. Doi: 10.13140/RG.2.2.30951.73125

[1] This section is based on a plenary talk [16] at the 29th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA) in Marburg. The PowerPoint slides we presented there can be downloaded (see reference section).

[2] This section is based on a plenary talk [19] we presented at the 30th IAFPA Annual Conference 2022 in Prague. The slides can be downloaded (see reference section).

[3] We are most grateful to Professor Bruce Lambert, director of the Center for Communication and Health at Northwestern University (Evanston, IL), for making his (updated) list of 1,250 confused medication brand names available to us.

[4] https://www.usp.org/products/usp-dictionary