DETERMINATION OF TRAINING SET SIZE FOR A STATISTICALLY-BASED WORDSPOTTER

M O'Kane, P Kenne and O White
Faculty of Information Sciences and Engineering
University of Canberra

ABSTRACT - Wordspotting in continuous speech is useful for automatically locating words for audio indexing purposes. Wordspotting is also the basic technology behind concept spotting, in which the location of enough members of a set of semantically-related words and phrases in a particular segment of speech is taken as an indication that the concept represented by that set is being discussed.

A set of experiments was conducted as a first attempt to determine the size of the database needed to train a statistically-based wordspotter. False negatives and the false positives are both treated as errors in wordspotting.

In the first experiment the size of the wordspotter training set needed was examined for the speech of a single speaker. Sufficient training data were collected until good wordspotting was achieved for this speaker. This experiment was then repeated for the speech of another speaker so that the variation of training set size as a function of speaker could be investigated. The training sets for the speakers were then pooled and the wordspotter was tested on test sentences for these speakers. The obvious generalisation experiment was then carried out in which the wordspotter was tested on test speakers who were not in the training set.

INTRODUCTION

A computationally efficient wordspotter was developed to perform at over 99% recognition in speaker-dependent mode. Results (less good) for the wordspotter working speaker-independently are also presented.

THE WORDSPOTTER

The wordspotter was constructed as follows. Twelve different 14-band broad encodings (O'Kane 1987) were computed from the fft of the input speaker. The bands for any one encoding are mutually exclusive. Each band receives one of the labels a,b,c,...,n. Speech which has the word to be spotted marked-up is then encoded using the twelve encodings. For each encoding a dictionary containing the encodings of the marked-up target words is constructed. An example of the dictionary for the first encoding for the word "crosstalk" occurring 119 times in continuous speech is given in Figure 1.

dfmnbcfn
dmnlbcfnan
nadmnabcbnan
naefmncb
nafemnb
nafmnabcn
nbgmnlcn
ncdcmnbnan
ncdmnacn
ncdncn
ncecnbcbn
ncemnabfdn
ncemncn
ncenmncn
ncfemnabnb
ncfmnbcbn
ncfmncbcn
ncfncbcn
ncmnbcbn
ncmncbn
ncnmncn
ndbmnbcn
ndclmnbcn
ndcnbcnan
ndedmnbcn
ndemlnb
ndemlncbn
ndemnbcbn
ndenmncbn
ndfcmlnbn
ndfdmnbcbn
ndfdmnbcn
ndfdmnbdcn
ndfdmnbn
ndfdnbdcn
ndfemnbc
ndfmncn

ndfnmnbcn
ndmlbc
ndmnabcn
ndmnbfn
ndmnkbn
ndmnlbcn
ndmnmbcnan
ndnbcbn
ndncbcn
ndnmlnbcbcbn
ndnmnbcn
ndnmncn
nedmncbcbcn
nedmncbcnan
nednbcn
nedncbn
nefmnbcn
nefnbcn
nemnbcbcn
nemncn
nenbcn
nenmnbcbnan
nenmncbcnan
nenmncn
nfdmnabcbn
nfegmnlbcbn
nfemlnbcn
nfemnkbdcn
nfemnkcbn
nfmnabcn
nfmnbcbn
nfmnbcnan
nfmnkb
nfncn

nfnmlncn
nfnmnbcbn
nfnmncn
nfmnbcn
nemnbcn
nfmnbn
ndfmnbcn
ndmnbn
ndmnbcn
nadmnb
nadmnbn
nbnmnbcn
ncmnbcn
ncmncn
ndcmnbc
ndfdmnabc
ndfmnbn
ndfmncbn
ndfnmncn
ndnbcn
ndnbn
nedmnbn
nenmncb
nfenbn
ncfnbcn
ncmnbn
ncncn
nefmn
nfnbcn
ncnmnbn
nemnbc
ndnmnbn
ndfmnb
ndnmn
ndmnbc
nmnbn
ndmnb

Figure 1: Encoding 1 dictionary for the word "crosstalk" derived from 119 examples of "crosstalk" in continuous speech from one female speaker

If one examines the twelve dictionaries one sees that the entries in any one dictionary are generally "close" in some word-nearness sense. This can be quantified by formally developing measures which reflect the number of letters by which any two dictionary entries differ. Or it can be quantified by considering all word pairs, triples, ..., n-tuples within the words of any dictionary. Table 1 shows the number of pairs to 6-tuples encountered for the dictionary in Figure 1. This table also gives the number of n-tuples theoretically possible for n=1 to n=6. The striking feature of these numbers is how small they are compared to the theoretical possibilities. That this holds for all twelve dictionaries can be seen in Table 2 which gives the 4-tuple results for all encodings.

n	2	3	4	5	6
number of n-tuples in dictionaries	55	131	199	238	239
number of n-tuples theoretically possible	182	2,366	30,758	399,854	5,198,102

Table 1: Number of different symbol paris, triples, ... 6-tuples in encoding 1 dictionary for "crosstalk" for one female speaker. Also shown is the number of n-tuples theoretically possible

encoding no.	1	2	3	4	5	6	7	8	9	10	11	12
number of 4-tuples in corresponding "crosstalk" dictionary	199	223	468	399	279	480	178	246	351	245	245	635

Table 2: Number of different 4-tuples in each of the twelve encodings dictionaries for "crosstalk" for one female speaker

This relatively low number of n-tuples is used in the wordspotter as follows. The utterance to be tested for the presence of the word to be spotted is encoded using the twelve encodings. Each utterance encoding is searched for the presence of the allowed 4-tuples for that encoding. Where two or more 4-tuples occur overlapping a potential find is marked. (Note that no word in the dictionaries is of length less than five.) Also marked are cases of (a 4-tuple + a letter + 2 or more overlapping 4-tuples) and (2 or more overlapping 4-tuples + a letter + a 4-tuple) and so on. This process is illustrated in Figure 2.

The potential finds for the twelve encodings are then OR-ed in time. Each potential find in each encoding is assigned a notional weight of 1. When two or more finds are OR-ed the weights are summed. After the OR-ing process is complete, all portions of speech which have a weight of 7 or more (it could be any number up to 12) are deemed to be the wordspotter's best attempt at the word being spotted. This wordspotter is simple to build and very efficient to run over test speech. All the work on separate encodings runs totally in parallel. Fast searching techniques ensure that the computation of the finds in each encoding are linear in the length of the test string. The OR-ing process is also computationally very fast particularly when run on special-purpose hardware.

It should be noted that the signal processing, the number of encodings, the number of bands in each encoding, the length of the n-tuple chosen and the OR-ed weight cutt-off point are all fairly arbitary.

```
n 0 8012 NO MATCH
b 8049 9129 NO MATCH
m 9129 10633 NO MATCH
b 10633 11730 NO MATCH
m 11730 13227 NO MATCH
k 13227 15580 NO MATCH
m 15580 16522 NO MATCH
a 16522 18641 NO MATCH
f 18641 19192 NO MATCH
g 19192 19912 NO MATCH
f 19912 20780 NO MATCH
d 21022 22375 NO MATCH
n 22412 27212 OCCURS          potential find of "crosstalk"
d 27226 29070 OCCURS
m 29070 30612 OCCURS
n 30612 32412 OCCURS
b 32423 33479 NO MATCH
c 33479 34993 NO MATCH
n 35012 38212 NO MATCH
f 38231 39012 NO MATCH
g 39012 40312 NO MATCH
f 40312 43951 NO MATCH
n 44012 47212 OCCURS          potential find of "crosstalk"
d 47301 48565 OCCURS
m 49012 50012 OCCURS
n 50012 51812 NO MATCH
b 51821 54000 NO MATCH
n 54012 55612 NO MATCH
b 55612 57012 NO MATCH
n 57012 59412 NO MATCH
c 59412 60603 NO MATCH
e 61039 61862 NO MATCH
k 61862 62412 NO MATCH
n 62412 66012 OCCURS
a 66012 66812 NO MATCH
b 67012 72012 NO MATCH
n 72012 84812 OCCURS
f 84851 85612 NO MATCH
n 85612 87012 NO MATCH
c 87021 88597 NO MATCH
a 89068 89720 NO MATCH
n 89812 91412 NO MATCH
c 91412 94530 NO MATCH
b 94530 94979 NO MATCH
```

Figure 2: Encoding 1 string of Speaker 1 saying "This is another crosstalk file. Crosstalk is the next word. I'm getting very.......". Potential finds (correct) for the word crosstalk are marked. Time shown in sample points. Sampling rate is 20 kHz.

"Intelligent" choices have been made for the purposes of this experiment. All these issues should be investigated for optimisation.

RESULTS

When the "crosstalk" wordspotter is built using continuous training speech from one female speaker containing 104 examples of the word "crosstalk" and tested on different continuous speech from the same speaker containing 56 examples of the word "corsstalk", the wordspotter finds all examples of the word "crosstalk" but chops the front and end of 8% of the examples.

When speech from this speaker containing 170 examples of "crosstalk" is used for the training phase and the wordspotter is tested on the same test speech, it works perfectly. When this word was tested on continuous speech containing 137 examples of crosstalk from Speaker 2 (male), 18 examples from Speaker 3 (female) and 41 examples from Speaker 4 (male) the results were as shown in Table 3.

	Wordspotting results
Speaker 1	100%
Speaker 2	15%
Speaker 3	39%
Speaker 4	73%

Table 3: Wordspotting percent results for Speaker 1-developed wordspotter tested on test speech from Speakers 1-4

A new wordspotter was created using continuous speech containing 118 instances of "crosstalk" from Speaker 2. Tested on other speech from Speaker 2 containing 19 examples of "crosstalk" this wordspotter achieved 84% correct.

When the two wordspotters were amalgamated and tested on all test speech from all speakers the results were as shown in Table 4.

	Wordspotting results
Speaker 1	98%
Speaker 2	89%
Speaker 3	61%
Speaker 4	85%

Table 4: Wordspotting results for (Speaker 1 + Speaker 2)-developed wordspotter tested on test speech from Speakers 1-4

The results for Speaker 1 drop slightly for the amalgamated wordspotter due to misfires. If the cut-off weight is adjusted to 9 then the results are 100% correct but results for the other speakers are then slightly lower. Perhaps the most surprising result is the result for Speaker 4 who does not sound at all like the training Speakers 1 and 2. Even the good result for Speaker 4 when using the Speaker 1 wordspotter alone is surprising as Speaker 1 is female and Speaker 4 is male.

CONCULSION

A computationally-efficient continuous speech wordspotter was constructed using speech from one speaker and then from two. It achieves good results on test speech from the training speakers and reasonable results on the two other speakers tested.

REFERENCES

O'Kane, M (1987) *Location and recognition of plosive consonants in continuous speech* ,Proceedings of the Eleventh International Congress of Phonetic Sciences, Tallin, Vol 2, 380-383.