A QUANTITATIVE METHOD FOR PERFORMANCE ANALYSIS OF AN ISOLATED WORD ASR SYSTEM

Michele Gubian[†]*, Luigi Arnone*[‡]*, Sergio Brofferio*[*]

[†]ICT, University of Trento, Italy - gubian@dit.unitn.it
[‡]AST labs, STMicroelectronics - Agrate Brianza - Italy - luigi.arnone@st.com
[*] DEI, Politecnico di Milano, Italy - brofferi@elet.polimi.it

ABSTRACT

This paper introduces a new measure of confusion between phones, based on isolated word recognition tests. This metric combines the advantages of previous measures, and excludes their disadvantages. It can be used for comparing the performance of two speech recognizers at phone level, providing a useful design tool. The main advantage is that tests are made on any set of recorded words, but measure of confusion is evaluated for a particular phone versus another one, and at the same time it is vocabulary independent. Note that manual phone segmentation is not needed. Furthermore, a suitable combination of several tests allows to obtain useful statistical paired tests. The advantages of this new method are illustrated on the basis of both artificial examples and comparisons between real ASR systems.

1. INTRODUCTION

In this paper we discuss the issue of finding a good metric for measuring confusion between phones in the context of Automatic Speech Recognition (ASR) research. In particular, we refer to single word speaker independent ASR systems [1] [2], based on phone units modelled by HMMs. In this kind of systems recognition is dictionary driven, i.e. a dictionary lists all words belonging to a given task, where each word corresponds to one or more phonetic transcriptions. Thus, a word model consists of a sequence of phone HMMs. During recognition all word models likelihoods are calculated, and the word giving the highest likelihood is taken to be the most probably spoken. To do this the well known Viterbi Algorithm [1] is used.

The importance of a metric for measuring confusion between phones relies upon the following aspects:

1. the knowledge of most confusable phones helps designer in managing resources (i.e. computation, memory, band);
2. assessing recognition performance at phone level makes performance evaluation independent of the task dictionary.

In order to explain the latter point, consider voice call in mobile phones. In this application, the dictionary is the user's address book, usually listing tens of names. Recognition performance strongly depends on the degree of phonetic similarity among those names. As an example, an address book listing just three names like: Maria (/m/a/r/i/a/), Marina (/m/a/r/i/n/a/) and Marisa (/m/a/r/i/z/a/)[1] results in a difficult recognition task. A metric for selectively measuring the degree of confusion between two specific phones can help preventing unexpected performance drops due to random word closeness.

This paper is organized as follows: section 2 describes four state of the art methods for measuring confusion between phones, showing the main drawbacks about each one; section 3 introduces a new method for obtaining such a measure; section 4 compares all of the above methods; section 5 describes the experimental framework used to validate the new method, and the validation itself; finally, section 6 draws some conclusions.

[1] all reported examples are based on Italian language; phonetic transcriptions are based on standard SAMPA for Italian [3]

2. MEASURING CONFUSION BETWEEN PHONES

This section illustrates four methods found in the literature for measuring confusion between phones, and analyzes pros and cons of each of them. Through the discussion we will refer to the following hypothetical situation:

- an isolated word HMM based ASR system has been trained, each HMM modelling a single phone;
- for illustrating purpose we will assume that the HMM related to the phone /n/ is "ill–modelled", yielding an intrinsic confusion between /n/ and /m/;
- all other HMMs are "well" trained.

We will use this fictitious example to analyze the capability of each method to detect the confusion /n/ $\rightarrow$ /m/.

2.1 Tests based on a generic dictionary

A set of recorded words is used as a test set, and a task dictionary including those words is used to form a word parallel grammar used during recognition. No particular constraint is taken into account in choosing words in the task dictionary. After the test phase, a phone Confusion Matrix (*CM*) is derived from recognition data. A phone *CM* is a double entry table accumulating recognition results at phone level: lines list correct (i.e. really uttered) phones and columns list recognized phones, such that the $CM(n,m)$ cell counts all the times phone /n/ has been mistaken for /m/.
Let's assume the word "fune" belongs to the test set, and the task dictionary includes both the words "fune" (/f/u/n/e/) and "fumi" (/f/u/m/i/). Two situations may occur:

1. the word "fune" is correctly recognized. This means that although /n/ is ill–modelled, the likelihood accumulated through the Viterbi path "fune" is higher than all the other likelihoods related to alternative word paths, in particular the one corresponding to the word "fumi". This results into increments of cells $CM(f,f)$, $CM(u,u)$, $CM(n,n)$, and $CM(e,e)$, thus nothing about the assumed /n/ $\rightarrow$ /m/ confusion is scored;
2. the word "fune" is mistaken for the similar one "fumi", due to the assumed /n/ $\rightarrow$ /m/ similarity. This results into increments of cells $CM(f,f)$, $CM(u,u)$, $CM(n,m)$ and $CM(e,i)$. Note that together with the expected raise of the value in $CM(n,m)$, another error is recorded in $CM(e,i)$. This is due to the fact that a word is mistaken as a whole, and this yields a sparse effect on *CM*.

Of course, other errors may occur, depending on what words actually compose the dictionary.
The main advantage in using this method for calculating *CM* cells is its simplicity. In particular, no special test set is needed and the whole *CM* matrix is calculated with a single test. The main drawback is that *CM* values *depend on the particular set of words included in the task dictionary*. An error may or may not occur depending on the phonetic distance among those words, and phone errors usually do not come out separately. In conclusion, this kind of test can be used for a quick evaluation of the most evident phonetic confusion trends, but is not a method suitable for obtaining selective and independent measures of phone recognition performance.

2.2 Tests based on manually segmented audio data

A straightforward way to avoid *CM* values dependence on the task dictionary is to cut audio files in segments, each one delimiting a single phone. After that, those audio segments are used as a test set. In this case task dictionary degenerates in a list of "words", each one composed by a single phone, and a phone parallel grammar is used during recognition. Of course, any dependence on a word dictionary is avoided. Two major drawbacks arise when employing this method:

1. segmentation is an expensive process. Even though partially automated via *forced alignment* [4], a manual check by a trained operator is needed [5] [6];
2. time boundaries of the employed phone segments are fixed. On the contrary, during recognition of a whole word those boundaries are determined by the recognizer, and vary according to its employed acoustic models. More clearly, the "correct" boundaries of a phone, derived segmenting by hand a particular audio track of a whole word, can be different from those assigned by the tested ASR system during recognition of that same audio track. This is especially true when poor modelling occurs, because the recognizer can make big mistakes in assigning those boundaries. These effects will not be detected if fixed length segments are used for testing.

2.3 Tests based on a phone dictionary

This kind of test employs a phone loop speech recognition grammar, letting the ASR system chain phones without constraints. In this case, as in the previous one, any dependence on a word dictionary is avoided. The main drawback here is in the kind of phone errors, which are different from those occurring when word recognition is driven by a regular word dictionary. These errors are insertions and deletions, especially of short phones like closures. As an example, suppose the word "fune" (/f/u/n/e/) is recognized as /f/u/cl/m/e/, where /cl/ is a closure phone. The dynamic programming (DP) algorithm used for classifying errors, based on the distance of Levenshtein between strings [7], assigns a substitution error /n/ $\rightarrow$ /cl/, and an insertion error of /m/, thus misleading the correct interpretation of the underlying problem of confusion /n/ $\rightarrow$ /m/, which was clearly the cause of the whole mistake.

2.4 Tests based on an ad hoc dictionary

This last case is a good compromise widely found in the literature [8]. Test set and task dictionary are made of very short and simple words, typically syllables, like /b/a/ , or bisyllables, like /a/b/a/. This choice fixes all drawbacks found in the previous three methods, but introduces two more ones:

1. tests cannot be made using any kind of recorded data. Even if many speech corpora include such kind of special words [8], an ad hoc test set has to be built if particular acoustic scenarios (e.g. car noise, noisy rooms) are to be tested, and this can be impossible in some circumstances (e.g. spontaneous speech);
2. simple syllables like /b/a/ are not a good sample of triphonetic contexts. For example, /t/ in /t/a/ is embedded in a quite simpler coarticulatory context than in the word "extra" (/e/cl/k/s/cl/t/r/a/).

3. A NEW METHOD

We propose a new method for calculating *CM* values, that overcomes all drawbacks mentioned above, and does not introduce any new one. This method is called *Minimal Pairs Synthetic Comparisons* (MPSC). We will illustrate this new method by an example: calculation of $CM(n,m)$.
A set of recorded words is used as a test set, with no phonetic constraints, just like in the case described in Section 2.1. Peculiar of this test is the task dictionary structure: it is composed of just *two* words, and *it depends on the input word*. In order to calculate $CM(n,m)$ we select from our audio test set all words containing at least one occurrence of the phone /n/. Let one of them be "fune". From its phonetic transcription, we derive a word parallel grammar composed of only two alternative paths:

- *fune*: /f/u/n/e/
- **fume*: /f/u/m/e/

The former word is simply the correct answer, and the latter is derived from the phonetic transcription of the former, substituting one occurrence of /n/ with /m/ – the only one in this case. The latter word, synthesized from the former, may not exist in the reference language vocabulary (Italian in this case), and this justifies the leading asterisk in "**fume*". Linguists call such a pair of words a *minimal pair* [9] [10], i.e. a pair of words differing in only one phone[2]. For each test utterance including at least an occurrence of /n/ a different corresponding grammar, similar to the one described above, is derived and it is used for recognition. Two counters, $CA(n,m)$ and $WA(n,m)$, accumulate correct and wrong answers, respectively. These counters hold all information needed, because the grammar structure allows just one kind of possible error: mistaking /n/ for /m/. After all utterances have been processed, $CM(n,m)$ is calculated simply by counting errors and normalizing with respect of the total number of performed tests:

$$CM(n,m) = \frac{WA(n,m)}{WA(n,m)+CA(n,m)}$$

This procedure can be applied for the calculation of an arbitrary cell $CM(p_1,p_2)$. Note that each cell value is calculated using a different and independent set of recognition experiments, and this independence can be exploited in order to obtain statistically significant performance measures, as will be clarified in section 5.

	2.1	2.2	2.3	2.4	MPSC
CM independent of task dictionary	No	Yes	Yes	Yes	Yes
Any kind of input	Yes	No	Yes	No	Yes
Single word working conditions	Yes	No	No	Yes	Yes
Selective measures	No	Yes	No	Yes	Yes
Minor error masking avoided	No	No	No	No	Yes

Table 1: Comparison of methods

4. COMPARISON OF METHODS

Table 1 compares the four state of the art methods described in section 2, labeled with their subsection numbers, and the new one proposed by us. Terms of comparison are: *CM* values independence of words in task dictionary; the possibility of making a test using any kind of stored speech data; whether test working conditions match real single word recognition ones; the possibility of restricting measures only on some phones versus some others; whether less likely phonetic mistakes can be measured. This last point is interesting, and is one of the items in favor of the MPSC method: the particular task dictionary structure, by means of which just one kind of phone confusion is measured at a time, makes it possible to measure confusion degrees between phonetically distant phones. In traditional tests this kind of measure is usually difficult and not reliable, because the choice of many alternatives in the task dictionary emphasizes the effects of most common errors *masking* the others, thus a few or no data at all is collected related to less likely mistakes. Experimental results exhibit this trend clearly, as will be shown in the next section.

[2]Linguists use the notion of minimal pair for defi ning the concept of phoneme from the more basic one of phone. Instead, we loosened the concept of minimal pair in order to describe our proposed method with a meaningful name. For further informations see [9] [10]

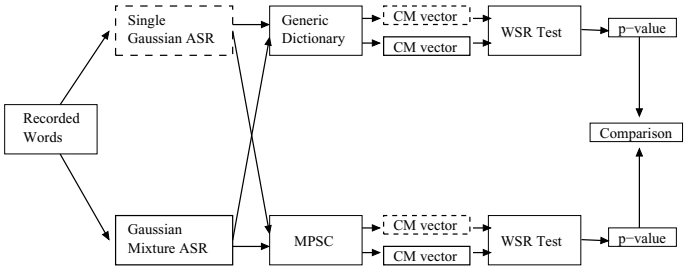

Figure 1: Structure of the experiments

Phone		*65 words*	*1000 words*	*MPSC*
Ph. set 1	ee	0.084	0.136	0.0457
Ph. set 1	EE	0.090	0.471	0.0004
Ph. set 1	e	0.185	0.363	0.0013
Ph. set 1	E	0.500	0.572	0.0001
Ph. set 2	ee	0.050	0.200	0.0478
Ph. set 2	EE	0.172	0.727	0.0011
Ph. set 2	e	1.000	0.144	0.0008
Ph. set 3	ee	0.090	0.133	0.0005
Ph. set 3	e	0.977	0.010	0.0013
Ph. set 4	e	0.500	0.117	0.0248
Ph. set 4	E	0.500	0.198	0.0045
Ph. set 5	e	0.500	0.071	0.0259

Table 2: p-values of WSR tests performed as in Figure 1

5. VALIDATION

We conceived a set of experiments in order to validate the new MPSC method for measuring confusion between phones. First of all we describe the experimental framework, then the structure of the experiments, partly inspired by [11] and [12], and finally we analyze the results.

5.1 Experimental framework

Speech data used in the experiments is part of a speech corpus called FONHABIT, a database of Italian isolated words spoken by different Italian native speakers and recorded in a quiet room. Recordings and database building and maintenance is carried out at AST labs, STMicroelectronics, Agrate Brianza (Italy). The FONHABIT speech corpus is based on a dictionary of Italian words, accurately selected in order to achieve two major goals:

- phonetic similarity among words
- diphone set coverage

while keeping the total number of words relatively small (less than 1000).

Ten different ASR systems based on phone units were built using HTK software [13]. All ASRs were trained using utterances of the same 65 random words spoken by 20 different speakers, 10 males and 10 females. Sampling rate is 10 KHz, and every 10 ms 39 MFCC components (12 cepstral coefficients and an energy parameter, plus their first and second order time derivatives) are extracted from a 25 ms frame in order to form observation vectors. Five different phone sets were employed, each one corresponding to a different vowel classification system. Each phone was modelled by a 3-states feed-forward HMM. Each of these five phone sets was used to build both a single Gaussian and a 3-components Gaussian mixture HMM based ASR, thus obtaining ten different ASR systems.

5.2 Structure of the experiments

Figure 1 shows the experimental scheme applied for comparing the new MPSC method to the state of the art method based on a generic dictionary (Section 2.1). Purpose of these experiments is to compare the detecting power of the different methods, measuring their effectiveness to detect performance differences between ASRs. The logic underlying this scheme is summarized below.

Two *CM* vectors are calculated using data coming from two different ASRs: one of these systems is designed to provide better recognition performance. The same *CM* vector pairs are calculated using both the traditional method (upper branch in Figure 1) and the new method (lower branch). Then, each of the two *CM* vector pairs becomes input of a hypothesis test for paired data, and *p-value* is extracted from each test. Finally, *p-values* are compared. The employed hypothesis test is Wilcoxon Signed Rank (WSR) test [14] for paired data, and test formulation is the following:

$$H_0 : CM_{\text{single Gaussians}} = CM_{\text{Gaussian mixtures}}$$

$$H_1 : CM_{\text{single Gaussians}} > CM_{\text{Gaussian mixtures}}$$

The alternative hypothesis H_1 expresses what we already know in advance, based on simple and reliable measures: an HMM based ASR system whose output densities are modelled with Gaussian mixtures performs generally better than a similar one whose output densities are modelled with single Gaussians. Exploiting this *a priori* knowledge we can measure how powerful each method is in detecting performance differences. Significance level of tests, inversely expressed by their *p-values*, is used to measure the capability of each *CM* calculation method in detecting differences in performance levels of different ASRs.

5.3 Experimental results

Tests were based on utterances of the same 65 words used in the training phase spoken by 6 speakers (3 males and 3 females) not included in the training set. We extracted the information related to the confusion of the phone /e/ versus all other vowel phones and some consonant phone.[3] More precisely: for each of the five phone sets mentioned in section 5.1, which differ only in the classification of vowels, we extracted all the possible *CM* values of the form $CM(e, vowel)$, where 'e' is a specific subclass of the phone /e/, and 'vowel' stands for all other phone vowels and some consonants. Thus for any fixed phone set and for any particular subclass of the phone /e/ we devised an experiment like the one described in Figure 1, extracting the specific $CM(e, vowel)$ vectors both from the single Gaussian and from the Gaussian mixture ASR based on the current phone set, and such calculations were performed using both the compared methods.

CM data extracted using the method based on a generic dictionary (Section 2.1, upper branch in Figure 1) were obtained employing two different modalities:

- using a 65 words parallel grammar, i.e. the same words selected for training and testing;
- using an extended 1000 words parallel grammar.

Table 2 summarizes our experimental results. The first column lists the specific phone set and 'e' subclass phone identifying the $CM(e, vowel)$ vector pairs; the different 'e' versions are expressed in an extended SAMPA notation[4] [3] [15]. The corresponding *p-values* derived from WSR tests performed on those vector pairs are listed in the remaining three columns: columns 2 and 3 list *p-values* calculated using the method described in Section 2.1, in the 65-words and 1000-words parallel grammar modalities mentioned

[3] The choice of /e/ is related to a forthcoming set of experiments aimed to quantitatively evaluate the benefits of employing distinct models for the Italian allophones e and E, i.e. closed and open /e/, in a given specific practical situation. Similar experiments will be conceived for the allophones o and O, i.e. closed and open /o/, and we intend to base our performance evaluation on MPSC method.

[4] The phone set 1 corresponds to the one used in [15]. The other four sets are obtained from the 1st by clustering some of the four subclasses in different ways. Note that the notation used is therefore not consistent across phone sets.

	65 words		*1000 words*		*MPSC*	
phone	single	mix	single	mix	single	mix
a	0	0	0.010	0.031	0.031	0.041
aa	0	0	0	0	0.010	0
e	0.021	0.010	0.082	0.072	0.186	0.134
ee	0.052	0.031	0.031	0.021	0.062	0.062
i	0	0	0	0	0.062	0.031
ii	0	0	0	0	0.031	0.010
j	0	0	0	0	0.031	0.010
l	0	0	0	0.010	0.031	0.010
n	0	0	0	0	0.062	0.021
o	0	0	0	0	0.041	0.010
oo	0	0	0.021	0	0.052	0.021
@sch	0	0	0	0.010	0.062	0.010
u	0	0	0	0	0.021	0.010
uu	0	0	0	0.010	0.041	0.031
w	0	0	0	0	0.021	0
E	0	0	0.206	0.103	0.278	0.175
O	0	0	0	0	0.010	0.010
OO	0.010	0	0	0	0.021	0

Table 3: Three couples of $CM(\texttt{EE}, vowel)$ vectors

above, and column 4 lists *p-values* calculated using the new MPSC method proposed by us.

In *all* cases *CM* vector couples calculated via MPSC method reached a better test significance level (i.e. a minor *p-value*) than the same vector couples calculated via a test based on a generic dictionary, showing that MPSC method is more suitable for determining differences in phone recognition performance between ASRs.

In order to supply a further insight of our experimental data we select one of the experiments from Table 2, namely the one summarized in the 2nd line, and we display all the *CM* vector couples which yielded the listed *p-values*. This specific experiment dealt with the measurement of the confusion between EE, i.e. open stressed /e/, versus other vowel phones, using the first of our five different phone sets. Table 3 shows these data. The first column lists all the phones which up to this point were always collectively labeled as 'vowel' in the expression $CM(e, vowel)$, and the following three couples of columns, having the same headings of Table 2, show the $CM(\texttt{EE}, vowel)$ actual (normalized) values which were input to the WSR tests.

The presence of many zeros in the first two couples of columns clearly shows the masking effect of less likely errors occurring in tests based on a generic dictionary (for example see the lines for i, ii and j). This effect disappears using the MPSC method, by use of which reliable and independent data can be obtained about confusion between any phone couple. This yields the much higher significance level achieved using our method.

6. CONCLUSIONS

The goal of the research presented in this paper was to find a good metric for determining phone recognition performance selective at phone level. After analyzing the state of the art, we proposed a new method, *Minimal Pairs Synthetic Comparisons* (MPSC), which revealed good characteristics in terms of feasibility, reliability and richness of information. The search of a reliable and inexpensive methodology for performance analysis was undertaken in order to create a base upon which devising automatic optimization flows and any sort of controlled experiments in the context of isolated word recognition and speech recognition in general. We showed that state of the art methods tend to be either unreliable or expensive; on the contrary we proved the goodness of our new method in terms of both reliability and cost.

From a theoretical point of view, we proved the goodness of our new method compared with the method based on a generic dictionary (section 2.1) in terms of statistical significance. We made no attempt to perform an exhaustive validation involving the other three described state of the art methods (sections 2.2, 2.3 and 2.4) because their drawbacks are mainly, but not only, related to cost, i.e. availability of suitable test audio data.

From a practical point of view, some further investigation is needed in order to clear some points related to how to apply the new MPSC methodology to triphones. In particular, whether phonetic contexts consistency has to be preserved or not in forming the alternative synthesized word in the two words parallel grammar (section 3) is still an open question. Similar adjustments are to be conceived in order to extend the use of our methodology to Continuous Speech Recognition (CSR) systems.

REFERENCES

[1] L. Rabiner, B. H. Juang "*Fundamentals of Speech Recognition*", Prentice Hall ed. , 1993

[2] B. Gold, N. Morgan "*Speech and Audio Signal Processing*", John Wiley & Sons ed. , 2000

[3] url: www.phon.ucl.ac.uk/home/sampa/italian.htm

[4] K. Sjölander "*Automatic alignment of phonetic segments*" - 2001 url: citeseer.ist.psu.edu/501039.html

[5] P. Cosi, D. Falavigna and M. Omologo, "*A Preliminary Statistical Evaluation of Manual and Automatic Segmentation Discrepancies*", Proceedings EUROSPEECH-91, 2ndEuropean Conference on Speech Technology, Genova, 24-26 September, 1991, pages 693-696. url: http://www.pd.istc.cnr.it/Papers/PieroCosi/cp-ES91.pdf

[6] P. Cosi, "*La Segmentazione delle Occlusive dell'Italiano Mediante Slam*", Atti del XXVI Cogresso Nazionale AIA, Torino,27-29 Maggio, 1998, pp. 311-316. url: www.csrf.pd.cnr.it/Biblos/piero-cosi.htm

[7] Sakoe, H. , Chiba, S. "*A similarity evaluation of speech patterns by dynamic programming*", Institute of Electronic Communications Engineering of Japan, July 1970, p.136

[8] ISOLET v1.3 – Center for Spoken Language Understanding @ OGI. url: www.cslu.ogi.edu/corpora/isolet/

[9] L. Serianni "*Grammatica Italiana – Italiano comune e lingua letteraria*" – UTET ed.

[10] A. Akmajian, R. Demers, R. Harnish "*Linguistica*" – Il Mulino ed.

[11] H. Strik , C. Cucchiarini, J. Kessens "*Comparing the recognition performance of CSRs: in search of an adequate metric and statistical significance test*", Proc ICSLP-2000, Beijin, 2000, pp. 740-743. url: citeseer.ist.psu.edu/401767.html

[12] H. Strik , C. Cucchiarini, J. Kessens "*Comparing the performance of two CSRs: how to determine the significance level of the differences*", Proc. of Eurospeech 2001, Aalborg, Denmark, Vol. 3, pp. 2091-2094. url: citeseer.ist.psu.edu/580851.html

[13] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland "*The HTK Book*", for HTK Version 3.2 - December 2002
url: htk.eng.cam.ac.uk

[14] M. Hollander, D. Wolfe "*Nonparametric Statistical Inference*", John Wiley & Sons ed. ,1973

[15] P. Cosi e J.P. Hosom ,"*High Performance "General Purpose" Phonetic Recognition for Italian*", Proceedings ICSLP-2000, International Conference on Spoken Language Processing, Beijing, Cina, 16-20 October, 2000, Vol. II, pp. 527-530. url: www.csrf.pd.cnr.it/Biblos/piero-cosi.htm