



Phonetic Decoder Assessment

C. Bourjot, A. Boyer and D. Fohr

CRIN/INRIA-Lorraine, BP 239, 54506 Vandœuvre cedex, France

Abstract : This paper describes a methodology to perform automatically a standard assessment of acoustic phonetic decoders. Results such as confusion matrix, deletion or insertion matrices are computed to provide a global overview of the system assessment. This method relies on the dynamic programming algorithm. It matches the outputs of the decoder with a standard transcription of the spoken sentence and its most common utterances derived with phonological rules. The errors made by the assessor are computed as errors of the decoder. In order to minimize this drawback, we guide the matching process by running an adaptation of the assessment algorithm to the decoder. The assessment of APHODEX, the acoustic phonetic decoder developed in CRIN, is performed and results are discussed.

1 Introduction

The present paper deals with a methodology to perform automatically a standard assessment of acoustic phonetic decoders. This work takes place in the ESPRIT European project SAM (System Assessment Method) which plans to define methodologies and to build common tools for the assessment of speech input/output systems.

Goals of acoustic phonetic decoders assessment are numerous; in order to improve their performances, the higher levels of continuous speech recognizers (lexical, syntactical,...) require often the main characteristics of the acoustic phonetic decoder level [1]. The knowledge of the kinds of errors made by the decoder is also useful to improve performances. Another aim of the evaluation is to provide a reliable means to compare different systems (this last point implies measure tools standardization).

Assessment is performed in two steps : the first step is just a computation of numerical results needed by the second one which deals with interpretation. We propose a method which provides a lot of numerical results (less or more detailed) in order to realize the previously mentioned goals. Then, we determine the confusion matrix for the units yielded by the decoder, an insertion matrix and a deletion matrix and some in-context deletion or insertion results. Other estimates like the overall recognition rate,

the unit recognition rate, the rate of insertion and the rate of deletion are also computed to give a global evaluation of the system. These rates are determined for all speakers and could also be computed for some categories of speakers (male/female) to precise the performance of the decoder. These results are obtained by running programs which ensure the statistical validity of the estimates. It is necessary to have big speech corpus and this implies at the assessment process to be completely automatic.

2 Automatic evaluation for acoustic phonetic decoder

A way to perform automatically acoustic phonetic decoder assessment consists in the matching the outputs of the decoder to the data presented for testing from a corpus of continuous speech. The alignment is generally performed by running a dynamic programming algorithm to maximize the similarity degree between the symbol sequence given by the decoder and a sequence which describes the test utterance [2]. Then the confusions, deletions and substitutions are determined. Two main problems occur :

- first, is dynamic programming well-adapted to the assessment process?
- secondly, what speech material is required?

Most of the dynamic programming algorithms used to match phonetic sequences have been conceived in the framework of automatic speech labelling [3] [4] or automatic speech recognition [5] [6]. The goals of these processes are different from those of assessment. It is effectively necessary to avoid that errors made by the alignment process are computed as errors of the phonetic acoustic decoder. Therefore it is important to find the most correct path. The dynamic programming algorithm gives the "best" path between both sequences of units according to a given criterion but not always the good one. As shown by Hunt [7], the dynamic programming algorithm tends to overestimate the recognition rate and its performances decrease as the error rate of the recognizer increases. Furthermore it was shown that it increases with the symbol sequence length. Then, the dynamic programming as usually described is not sufficient and it is necessary to introduce in the matching process informations about the main characteristics of the decoder. It is the solution we chose to implement.

The second question we have to answer about the speech material is related to the previous one. It is possible to have more or less information about what has been pronounced, facilitating the matching algorithm. These information can be :

- a standard transcription of the spoken sentence obtained by a text to phoneme translation or by a phonetician,
- a perceptual transcription of the sentence which requires a listening of the speech signal,
- a perceptual transcription with a segmentation and a labelling of the signal,
- a synthesis of a given sequence of phonemes [8].
- a semi automatic labelling of the speech signal.

The third point is certainly true to reality and does not require a sophisticated matching algorithm. But as previously mentioned, a big corpus is required to give statistically reliable results [9] [10]. It is the reason why the points 2 and 3 are time consuming and the constraint of homogeneity is hard to respect because of the number of labellers which are involved in this task. The point 4 requires a good synthesizer and cannot take into account all the speaker characteristics. The point 5 requires a standard or by hearing transcription and a human correction which is fastidious.

Thus we chose to use a standard phonemic transcription of the corpus. The main difficulty is to consider with homogeneity the different pronunciations of a sentence [11]. We use phonological rules to derive the most common pronunciations from the standard transcription (possible utterances of a word, liaisons between connected words,...).

These rules allow to improve the alignment algorithm and avoid taking into account the variants of a sentence as errors of the phonetic decoder. Nevertheless, some pronunciations are not foreseen and will be considered as errors of the decoder but in the same way for all systems.

3 Method

As shown in the previous paragraph, our assessment algorithm relies on a dynamic programming algorithm adapted to the decoder.

This is to minimize the errors made by the alignment process which are computed as errors of the system. The assessment is performed in two main steps : a step of adaptation which computes intermediate results which will be used during the second step to help the dynamic programming algorithm to determine the correct warping path.

3.1 Adaptation

The step of adaptation is performed as follow : it determines automatically a coarse estimation of the most common errors made by the decoder. A confusion matrix, called an adapted matrix, an insertion rate and a deletion rate for each unit of the decoder are estimated in first approximation.

To compute these intermediate results, a dynamic programming algorithm is applied to a small continuous speech hand labelled corpus. This latter must contain about 1000 phonemes to be representative enough of the behavior of the decoder. It must be multispeaker in order not to reflect the characteristics of one particular speaker. Of course, both sexes must be present. The boundaries are used during the matching process to guide the dynamic programming algorithm. We do the following hypothesis :

- a phoneme must be confused with only one other phoneme
- a phoneme could be associated with another one only if there is a compatibility of their boundaries : either inclusion of one in the other one or only a little gap separates them.

We use the basic local constraint with the weighting coefficients (1.1.1). The recursive formula we use is corrected by the factor ddi in case of confusion where $ddi=infinite$ if there is no intersection between both segments, a factor proportional to the size of the intersection otherwise.

A basic confusion matrix is required to compute the local distance during the matching process. Its coefficients are :

- $1 - (n - 1) \times \epsilon$ if both units are identical (n is the number of units yielded by the decoder and ϵ a very small real),
- ϵ otherwise.

The main advantage of this matrix is that it is very simple to build and does not require any knowledge about the units.

3.2 Evaluation

The dynamic programming algorithm used in this step is quite similar to the previous one. The local constraint is the same. We maximize the probability of coincidence between both symbol strings. We compute the cumulated product of the probability of matching two phonemes. This basic probability is issued from the previous adapted matrices and is determined as the frequency of insertion (or deletion or confusion) of the considered unit during the adaptation phase. As previously mentioned, phonological rules derive the main pronunciations of the spoken sentence from the standard transcription. All the derived sentences are compared by dynamic programming with the outputs of the decoder. The utterance which realizes the best score is considered as the good one, and is used to compute the assessment result. At present, only three rules are used. This algorithm must run on a big enough corpus in order to provide statistically significant results.

4 Experiment

This assessment method has been used to assess APHODEX, the acoustic decoder which is developed in CRIN. In the first subsection, a quick description of the system is given. The following subsection provides a description of the corpora used to perform evaluation. Then, results obtained are given and the main conclusions are summarized.

4.1 Description of the APHODEX system

APHODEX is an acoustic phonetic decoder integrating the phonetician's competence into a knowledge based system. This system consists of :

- a set of pre-processors that operate on the speech wave and perform a coarse segmentation as well as acoustic analysis of various types on each segment.
- an expert system that labels segments and, if need be, refines the coarse segmentation provided by the pre-processors.

More information about this decoder can be found in [12] [13]. The first step of the decoding process identifies the macro-classes : fricative (FR), plosive (PL), vocalic nucleus (VN), sonorant (SN), fricative vowel (FV). The second step starts reasoning with the information collected during the previous step and identifies the best candidate (s) by running expert rules.

At present time, we only do the assessment of the macro-classes.

4.2 Corpora

The corpus consists of 11 speakers (6 male, 5 female) speaking 'La bise et le soleil'. It has been made in the framework of the GRECO Communication Parlee. All this speech signal has been labelled by an expert phonetician and a standard transcription is available. It contains about 4500 phonemes.

Three speakers (2 male, 1 female) are used for the adaptation step. This part of the corpus represents about 1200 phonemes.

During the second step, all the speakers are used.

4.3 Phonological rules

Some phonological rules are used. For example :

- pause can be inserted between 2 words,
- schwa can be pronounced or no,
- insertion of a nasal consonant may occur after a nasal vowel if it is followed by a plosive.

4.4 Adaptation results

The figure 1 indicates the adapted confusion matrix. The phonetic alphabet is SAMPA. The number of deletions and the recognition rate are precised for each phoneme and the rate of insertion for each classe of phonemes. For example, phonemes R and l are often deleted, and the analysis of their contexts show that they are deleted after a pause in half cases. This rule will be used in the evaluation process to guide the matching process. Another meaningful indication is the maximum number of consecutive deletions and insertions. These data will determine the weighting coefficients of the local constraint.

4.5 Evaluation results

Figure 2 shows the performances and the limits of the evaluation process. The first one illustrates the use of the rule about "l" determined during the adaptation. The second one shows that the assessor hypothesizes that the segment SN has been inserted after VN. In fact, no insertion has been committed, because the standard transcription does not reflect what has really been pronounced. The speaker says "le" instead of "un". The line "hand labelling" is only indicated to give an idea of what has actually been uttered. It is not used by the evaluation process.

5 Conclusion

This paper gives a method to perform automatically the acoustic phonetic decoder assessment.

This method has some drawbacks such as :

- the time required to realize the assessment of the decoder increases fastly with the amount of phonological rules.
- some errors are made by the matching algorithm. Even if information about the decoder are provided, it is impossible to avoid some mismatching. Nevertheless, we would like to mention that human experts make errors too when they do not have the signal segmented.

The main advantages of this method are :

- it does not require a hand labelling of the whole corpus.
- the matching process is guided by information on the decoder which are collected during the adaptation phase. Thus, the most common errors made by the system are known and this knowledge helps the algorithm to find the "good" warping path.
- the process runs completely automatically and requires no manual intervention. Then the documentation about the way tests are driven is standard. It facilitates the interpretation and makes comparable the results obtained [14].
- this method is easy to adapt to any kind of language and any kind of units yielded by the decoder.

6 Bibliography

- [1] Bourjot C., Boyer A., Mari J.-F., 'Methodology about assessment of large vocabulary systems', 7th FASE Symposium, book 1, pp161-169, EDINBURGH 1988.
- [2] Picone J., Goudie-Marshall K., Doddington G., Fisher W., "Automatic text alignment for speech system evaluation", ICASSP 86, pp780-784, Tokyo, 1986.
- [3] Wagner M., "Automatic labelling of continuous speech with a given phonetic transcription using dynamic programming algorithm", ICASSP 81, pp1156-1159, 1981.

[4] Leung H.C., Zue V.W., "A procedure for automatic alignment of phonetic transcription with continuous speech", ICASSP 84, pp271-274, San Diego, 1984.

[5] Smith A.R., Sambur M.R., "Hypothesizing and verifying words for speech recognition", Trends in Speech Recognition, Lea W. A. editor.

[6] Haton J.-P., "Une methode dynamique de comparaison de chaine de symboles de longueurs differentes : application à la recherche lexicale", CRAS serie A278, pp1525-1530.

[7] Hunt M.J., "Evaluating the performance of connected word speech recognition system", ICASSP 88, pp457-460, New York, 1988.

[8] Lennig M., "Automatic alignment of natural speech with a corresponding transcription", Speech Commu-

[9] Montacie C., Chollet G., "Systeme de reference pour l'evaluation d'application et la caracterisation de base de donnees pour la reconnaissance de la parole", 16eme JEP, pp323-326, Hammamet, 1987.

[10] Tubach J.-P., "Problemes et methodes en evaluation de la reconnaissance phonetique", JEP 84, pp109-110, Bruxelles, 1984.

[11] Pister-Bourjot C., Haton J.-P., "Automatic learning : an approach to the adaptation of a speech recognition system to one or several speakers", Speech Communication 1987, Elsener Science Publisher, North Holland, 1987.

[12] Fohr D., "APHODEX : un systeme expert en decodage acoustico-phonetique de la parole continue", these de doctorat de l'Universite de Nancy 1, 1986.

[13] Carbonnel N., Fohr D., Haton J.-P., "APHODEX, an acoustic phonetic decoding expert system", International Journal of Pattern Recognition and Artificial Intelligence, Vol. 1 no-2, pp207-222, 1987.

[14] Pallet D.S., "Performance assessment of automatic speech recognizers", Journal of research of the national bureau of standards, vol-90, number 5, september-october 85, pp371-387.

	nb	PL	VN	FR	SN	FV	del	rec.rate
p	32	31	0	0	0	0	1	(96 %)
b	22	13	1	0	2	0	6	(59 %)
t	57	55	0	0	2	0	0	(96 %)
d	28	12	0	0	5	0	11	(42 %)
k	46	38	1	1	0	0	6	(82 %)
g	2	0	0	0	1	0	1	(0 %)
f	30	0	0	10	14	0	6	(33 %)
v	24	5	1	0	8	0	10	(20 %)
s	6	0	0	6	0	0	0	(100 %)
z	12	0	0	11	0	1	0	(100 %)
ʃ	74	0	0	70	2	0	2	(94 %)
ʒ	20	0	0	14	3	2	1	(80 %)
m	36	2	0	0	28	0	6	(77 %)
n	9	0	0	0	9	0	0	(100 %)
Ń	0	0	0	0	0	0	0	(-)
J	30	0	0	3	14	0	13	(46 %)
w	12	0	0	0	9	0	3	(75 %)
R	81	2	2	0	47	0	30	(58 %)
l	105	1	1	0	67	0	36	(63 %)
H	9	0	0	0	3	0	6	(33 %)
o~	24	0	22	0	0	0	2	(91 %)
a~	33	0	32	0	0	0	1	(96 %)
e~	6	0	6	0	0	0	0	(100 %)
oe~	9	0	9	0	0	0	0	(100 %)
i	52	0	44	0	0	3	5	(90 %)
e	54	0	44	0	1	1	8	(83 %)
ai	58	0	55	0	0	0	3	(94 %)
a	90	0	80	2	2	0	6	(88 %)
O	40	0	36	0	0	0	4	(90 %)
o	38	0	28	0	0	0	10	(73 %)
u	15	0	14	0	1	0	0	(93 %)
y	27	0	23	0	1	0	3	(85 %)
z	3	0	3	0	0	0	0	(100 %)
9	12	0	12	0	0	0	0	(100 %)

	uttered	found	inserted
PL	280	241 (86%)	10 (4%)
FR	142	114 (80%)	8 (6%)
VN	538	484 (90%)	20 (4%)
SN	282	177 (63%)	63 (22%)

maximum consecutive deletions : 2
maximum consecutive insertions : 2

FIGURE 1 : RESULTS OF THE ADAPTATION

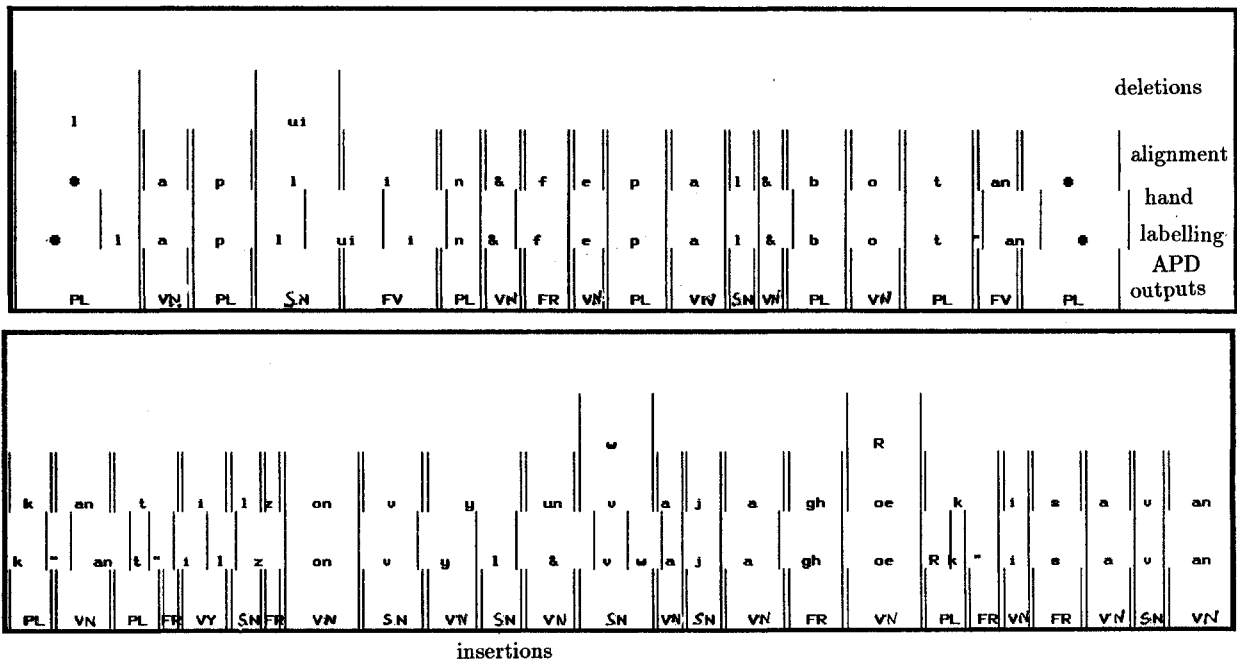


FIGURE 2 : EXAMPLES OF ALIGNMENT PROCESS