USEFULNESS OF PHONETIC PARAMETERS IN A REJECTION PROCEDURE OF AN HMM BASED SPEECH RECOGNITION SYSTEM

K. Bartkova & D. Jouvet
France Télécom, CNET /DIH/RCP, Technopole Anticipa, 2, Avenue P. Marzin,
22307 Lannion Cedex, France
Tel: (33)-2-96-05-10-58, e-mail bartkova@lannion.cnet.fr
Fax: (33)-2-96-05-35-30

ABSTRACT

The aim of this paper was to study the efficiency of sound duration, degree of sound voicing and sound energy in a rejection procedure of an automatic speech recognition system. A modelling of the three parameters was achieved using statistical models estimated on vocabulary words, out-of-vocabulary words and noise tokens. The rejection of out-of-vocabulary words and noises depended on the score obtained by comparing the probability given by the different models. However, such an approach also cause false rejection (rejection of vocabulary words). A trade-off was therefore necessary between the false rejection rate and the false alarm rate on out-of-vocabulary words and noise tokens. The degree of voicing turned out to be the most efficient parameter for rejecting noise tokens ; it reduced the HMM false acceptance rate from 6.3% down to 2.3% for the same amount of false rejection rate (9%). The duration parameter provided better performance for laboratory data, reducing the error rate on French numbers from 3.1% to 1.5% for a 5% false rejection rate.

1. INTRODUCTION

Some researchers elaborated sophisticated approaches to capture phonetic parameter variations such as the sound duration or F0 curve movement, in order to help speech recognition systems [1]. However models were often trained on hand-segmented data and their efficiency in term of recognition error rate reduction was not evaluated [2]. Some models were dedicated to the segmentation task of the speech signal (sentence boundary detection, stress detection...) without switching its output to a speech recognition system [3].

Some studies tried to use phonetic knowledge to constrain speech recognition systems. The phonetic parameters integrated so far into speech recognition systems were often modelled in a quite simple manner such as minimal sound duration [4]. Models were also used on isolated word or connected word corpora to rescore the N-best solutions obtained by the HMM [5]. However simple, these approaches allowed to evaluate to what extent non spectral parameters could improve speech recognition performance no matter whether the segmentation of the speech signal yielded by the system is correct or not, or whether the detection of the parameter (F0, intensity...) is accurate or not.

The aim of our study was to use phonetic parameters in automatic speech recognition in order to improve out-of-vocabulary words and noise tokens rejection when speech recognition systems are used in Interactive Vocal Services (IVS). When only keywords are necessary to successfully use a vocal service, garbage models are trained to capture and reject the non-keywords as well as different kinds of noises produced by the user's background. Errors committed by the system have various impacts on the service information delivery. For instance a false alarm (incorrect word acceptance) will be less tolerated by the user than keyword rejection. Indeed, false alarms and keyword substitutions lead to access non-required information while a keyword rejection simply forces the user to repeat the word. For that reason, a great effort is devoted to reduce false alarm rates and to make the IVS services more user-friendly [6]. However, a too high level of keyword rejection would not be tolerated by IVS users either. Therefore, in this study, the efficiency of the parameter modelling was measured by the trade-off between keyword false rejection rate and non keyword acceptance rate.

Three phonetic parameters are evaluated in this paper : sound duration, degree of sound voicing and sound energy. These parameters were used in the post-processing procedure of an automatic speech recognition system. After a first recognition pass using spectral information (mel frequency cepstral coefficients) it may be useful to apply different kinds of knowledge in order to accept or reject this first « spectral » selection.

2. DATABASE AND SYSTEM OVERVIEW

Sound duration, sound energy and sound voicing degree were modelled and evaluated using field data (data collected from IVS in operation) and laboratory data.

The laboratory corpus contained French numbers between 00 and 99 and the field corpus from the « Baladins » application [6], 26 French keywords. As the laboratory data was recorded under supervised condition, it did not contain tokens of noise or out-of-vocabulary words while the field corpus contained both of them. Both corpora were recorded over the telephone network by several hundreds of speakers. The corpora were split up into two equal parts, one part was used for training the phonetic model parameters and the other part for testing their efficiency for speech recognition.

The recognition system used in this study was the CNET system PHIL90 and the HMM acoustic modelling units were allophones (context-dependent modelling of phonemes) [7].

3. PARAMETER MODELLING

The aim of parameter modelling is to compute a likelihood ratio allowing to decide if a word has been correctly recognised or not. Therefore, at least two models had to be set up: one for modelling the phonetic parameters corresponding to correctly recognised word, and one for incorrectly recognised tokens (recognition errors and false alarms). Each phonetic model is represented by a Gaussian density, and defined by its mean value and standard deviation parameters.

Phonetic parameter models were estimated for every phoneme building up a particular word. The reason for the word-dependent approach was the following: HMM segmentation is quite poor and when a phoneme occurs only in few contexts, its boundaries are not always phonetically correct. However, it appeared that segmentation errors between phonemes were produced fairly consistently: this means that a segmentation error is often the same all along the corpus. Consequently, as the number of words in the vocabulary was not too high, a word and context-dependent modelling of the phonetic parameters seemed to be the most appropriate for these corpora. This modelling can be considered as a triphone model albeit each triphone is connected to a particular word and also to a particular position in the word. No parameter sharing was carried out among triphones. In a similar way an extra model was estimated for each word as a whole unit.

As laboratory data contained only vocabulary words, the two models were associated to the phonetic parameters for correctly recognised vocabulary words on the one hand and an for incorrectly recognised vocabulary words on the other hand.

Field corpus contained not only vocabulary words but also out-of-vocabulary words and noise tokens. Two approaches were therefore experienced. The first approach consisted in modelling all the incorrect answers (substitution errors, false alarms on out-of-vocabulary words and noise tokens) with one single model. In the second approach three different models were calculated for incorrect answers: one for incorrectly recognised vocabulary words, one for false alarms on out-of-vocabulary words and finally one for false alarms on noise tokens. It turned out that breaking down one incorrect model into three sub-models allowed us to take into account more accurately the differences existing among the three classes of bad tokens. The performances obtained with three sub-models where better than those obtained by the one single model.

Figure 1 gives an example of voicing degree modelling for a particular word of the field corpus (« Lannion »). For each unit and each model the mean value and the standard deviation are represented. The first value was yielded by the correctly recognised words modelling, the second one by the keyword substitution error modelling, the third one by out-of-vocabulary word false alarm modelling and finally the last one by noise token false alarm modelling. Besides the phoneme units, the word unit is also represented on the diagram.

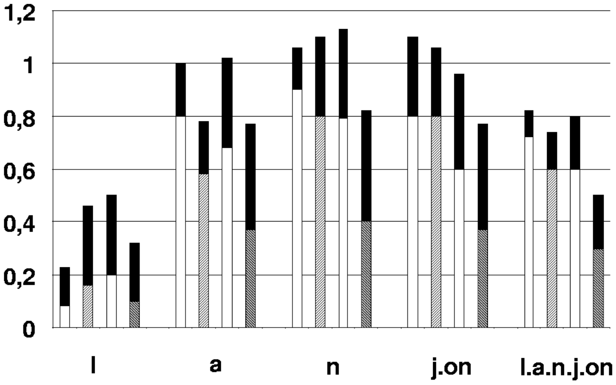

Figure 1 : Voicing degree modelling for the word « Lannion ».

3.1. Duration model

The sound duration was normalised by the duration of the word in which the sound occurred. In each phoneme succession, where a sound segmentation was considered as being difficult, phoneme clustering was carried out. For instance, instead of having two duration values for a succession of phonemes such as the semi-vowel [j] followed by the nasal vowel [on] in the word [l.a.n.j.on], only one duration was modelled for the sequence [j.on]. In this way, the correct model is supposed to present less variation (implying lower standard deviations) in the pronunciations of the different speakers, making it easier to separate the correct values from those associated to incorrect events.

In the correct duration modelling, the variations in the duration observed can be partly explained by phonological rules, partly by the speech processing technique. Thus, it was observed that the significant variation of the normalised sound duration occurred for word initial phonemes and for word final syllables. The duration variation of the word initial phoneme may be induced by the speech recognition technique. If speech is detected too early or too late by the noise-speech separation algorithm, phoneme standard deviation value may be affected. On the contrary, the final syllable variation can be due to the phonological rule of the French language : the French accent falls on the last syllable of the stress group (here last syllable of the word) and its realisation gives a relative freedom to speakers as far as sound duration is concerned.

3.2. Sound energy model

Since energy is highly variable, several normalisation procedures were developed when modelling sound energy values: normalisation by the whole word energy, by the highest vowel energy of the word, by all the vowels composing the word and finally by the middle part (three middle frames) of the vowels composing the word. The drawback of vowel energy normalisation was that in monosyllabic words the normalised vowel energy became equal to 0. Therefore, in monosyllabic words, consonants remained the only elements to provide information about correct or incorrect energy values. The best results, as reported in chapter 5, were obtained after normalisation by the energy of all the vowels building up the word.

3.3. Sound voicing degree model

The sound voicing degree was obtained as the ratio between the number of voiced frames and the total number of frames of each sound. Thus, an utterly voiced phoneme had a value of 1 and an utterly devoiced phoneme was equal to 0. Nevertheless, it scarcely happened that a phonologically voiced phoneme had a voicing degree equal to 1 or a phonologically unvoiced phoneme a voice degree equal to 0. In fact our data were recorded through the telephone line, hence the voicing detection was often perturbed by additional noises. I addition, as the sound segmentation was automatically performed by the HMM models, the sound boundaries were not always correctly detected, hence a voiced sound could be corrupted by an unvoiced context and vice versa.

Generally speaking, the results showed, that vowels had a higher voicing degree than voiced consonants. Among the vowels, nasal vowels had a higher voicing degree than oral ones and vowels in voiced contexts had a higher voicing degree than vowels in unvoiced contexts (silence was considered as an unvoiced context).

4. PHONETIC SCORE

A phonetic score was calculated to reject or to confirm the best solution yielded by the HMM recognition system.

The phonetic score was calculated from the likelihood ratios associated to each phonetic parameter. For a given answer, the logarithm of the likelihood ratios were summed up over the phoneme and global models associated to the recognised word. If this phonetic score was inferior to an a priori chosen threshold value, the answer was rejected, otherwise it was considered as a correct one. Unfortunately, the phonetic score of a correctly recognised word could also be below the threshold. Therefore a trade-off had to be worked out between the number of false acceptances and the number of false rejections (vocabulary word rejection rate).

In a previous paper [5], the efficiency of the duration and voicing degree was evaluated for rescoring the N-best candidates, proposed by the HMM. Contrary to the present paper, the phonetic score was recombined with the HMM score and a mixed HMM-phonetic score was used in the post-processing. Using a rule-based sound duration modelling, the reduction of the recognition error rate was about 7% on three isolated word French corpora recorded over the telephone network by several hundred of speakers.

A rule-based sound voicing model was also previously tested for rescoring the N-best candidates [8]. This parameter turned out to be efficient for field data and allowed to reduce the HMM error rate by 21%.

The previously quoted studies proved that phonetic parameters can also be efficient in recognition error recovery. However the HMM substitution rate of the vocabulary words is low in comparison with a relatively high false acceptance rate (acceptance of out-of-vocabulary words and noise tokens). Therefore, in the present paper, the phonetic score alone (without combining it with HMM score) is only applied to the first solution yielded by the HMM.

5. EXPERIMENTS

As PHIL90 contains efficient garbage modelling trained on out-of-vocabulary words, it seemed important to compare the results obtained using phonetic scores with those obtained using the HMM garbage models.

As shown in Figure 2, the voicing degree is particularly well adapted to noise token rejection. For 9 % of false rejection rate, the HMM alone leads to a false alarm rate of 6.3% (on noise tokens). Using the voicing degree reduces this false alarm rate down to 2.3%. As far as out-of-vocabulary words are concerned, using phonetic parameters did not improve their rejection rate. In fact, the HMM rejection performance and the phonetic rejection performance remained very similar.

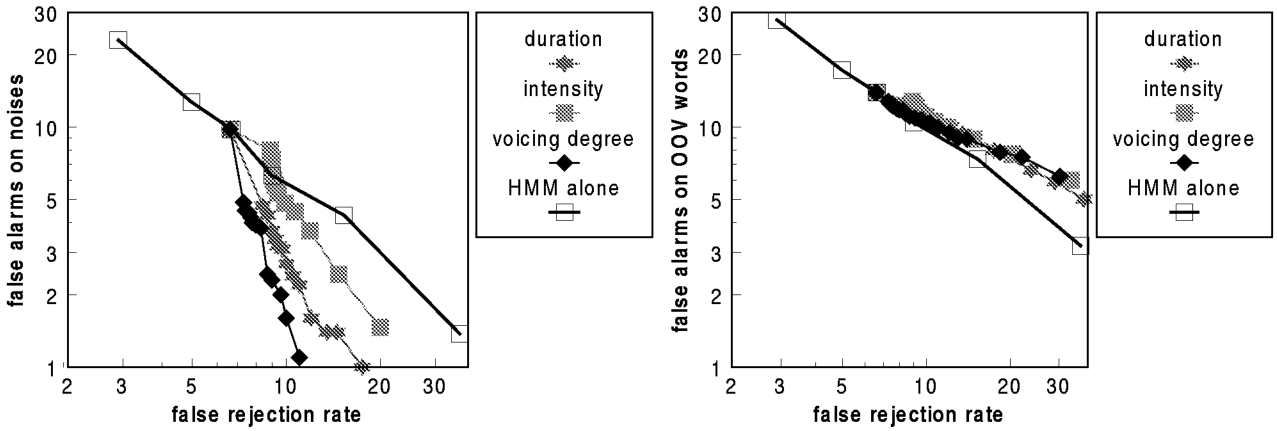

Figure 2: Comparison of rejection procedures on out-of-vocabulary words (right) and noises (left).

Figure 3 represents the evolution of the substitution and false rejection rates on the French number laboratory data. The HMM error rate was 3.1 %. This rate was the starting point for applying the phonetic parameters. The curves were obtained by modelling the three phonetic parameters separately and jointly (i.e. the three scores were acting together). A joint application of the 3 parameters did not yield better results than applying each parameter separately. In this application, the best results were obtained using the duration model which, for example, for a false rejection rate of 5% leads to a 50% reduction in the substitution error rate.

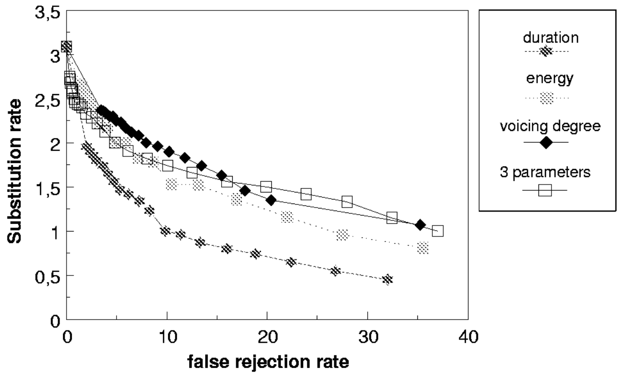

Figure 3: Substitution and false rejection rates on laboratory data.

6. CONCLUSION

This study was dedicated to phonetic parameter modelling and the use of these parameters in an automatic speech recognition procedure. Parameters where used in a post-processing procedure in which a phonetic score was computed to reject or confirm the best HMM solution.

It was shown that phonetic parameters can be efficiently used to improve the performance of automatic recognition systems. One of the parameters, the sound voicing degree, turned out to be very efficient for noise token rejection. A second parameter, the sound duration, worked well on the corpus of French numbers probably as it contained quite significant word length variations.

In this study, the units used for the phonetic parameter modelling were word-dependent triphones, and no parameter sharing was performed among triphones. However, such a parameter modelling is only suitable for small vocabularies. For large vocabulary processing, it appears necessary to use a vocabulary-independent modelling based on phonetic rules.

REFERENCES

[1] **Ostendorf M. and Ross K.:** « A Multi-level Model for Intonation Labels », Computing Prosody, pp. 291-308, 1996.

[2] **Pols L. and al.:** « Modelling of phone duration (using the TIMIT database) and its potential benefit for ASR », *Speech Communication* 19, pp:161-176, 1996.

[3] **Nakai M. and al.:** « Accent Phrase Segmentation by F0 Clustering Using Superpositional Modelling », Computing Prosody, pp. 343-359, 1996 .

[4] **Gupta V. and al.**: « Use of minimum duration and energy contour for phonemes to improve large vocabulary isolated-word recognition », *Computer Speech and Language,* vol. 6, pp.345-359, 1993.

[5] **Bartkova K. and al.:** *« Using Segmental Duration Prediction for Rescoring the N-best Solution in Speech Recognition »,* ICPhS95, Stockholm, Vol.4, pp:248-251, 1995.

[6] **Sorin C. and al.:** « Operational and experimental French telecommunication services using CNET speech recognition and text-to-speech synthesis », *Speech Communication* 17, pp:273-286, 1995.

[7] **Jouvet D. and al**: *« On the modelisation of allophones in an HMM based speech recognition system »,* EUROSPEECH, Genova, pp. 923-926, 1991.

[8] **Bartkova K. and al.:** *« Introduction de paramètres phonétiques en post traitement d'un système markovien de reconnaissace de la parole »,* JEP Avignon, pp. 305-308, 1996**.**