An elitist approach for extracting automatically well-realized speech sounds with high confidence

Jean-Baptiste Maj, Anne Bonneau, Dominique Fohr, Yves Laprie

An elitist approach for extracting automatically well-realized speech sounds with high confidence

Jean-Baptiste Maj, Anne Bonneau, Dominique Fohr
Yves Laprie

Speech Group
Loria-INRIA / CNRS, Nancy, France
jbmaj@loria.fr

Abstract

This paper presents an 'elitist approach' for extracting automatically well-realized speech sounds with high confidence. The elitist approach uses a speech recognition system based on Hidden Markov Models (HMM). The HMM are trained on speech sounds which are systematically well-detected in an iterative procedure. The results show that, by using the HMM models defined in the training phase, the speech recognizer detects reliably specific speech sounds with a small rate of errors.

1. Introduction

The present article introduces a method called 'elitist approach' for extracting automatically well-realized speech sounds with high confidence. By well-realized, we mean that the speech sounds have well marked phonetic features. From an acoustical and perceptual point of view, a same sound in the same phonetic context has more or less marked acoustics cues and a highly variable level of intelligibility. Considering that well marked phonetic features should be identified very reliably in automatic speech recognition we defined a series of cues, called 'strong cues', specially designed to identify this kind of feature [1].

These cues were defined from acoustic-phonetic knowledge and tested by means of a semi-automatic rule based speech recognition system. Since the purpose of 'strong cues' is to find out features that are well-realized from an acoustical point of view and make no error, they are not triggered systematically.

A very reliable detection of well-realized sounds may lead to two kinds of application : provide an automatic speech recognition system with reliable information on the one hand, and improve the intelligibility of speech through the enhancement of well-realized sounds on the other hand.

In order to detect 'well-realized' sounds in a fully automatic manner, we design in the present article, an elitist learning of HMM that make very reliable sound models emerge. The learning is iterated by feeding sounds identified correctly at the previous iteration in to the learning algorithm.

The elitist approach is based on a speech recognizer and a DTW algorithm. The speech recognition system works with Hidden Markov Models (HMM) and a learning stage is performed to train these models. The DTW algorithm compares the output of the speech recognizer and the phonetic annotation of corpora for identifying well- and wrongly-detected speech sounds. The phonetic annotations of the corpora are changed according to the results of the DTW algorithm, and an iterative procedure is carried out to improve the accuracy of the HMM-based phoneme models. In this study, the well-detected phonemes are assumed to be well-realized speech sounds.

A similar approach was proposed by Schwenk [2] to guide the learning of multilayer perceptrons in the context of automatic speech recognition. This approach is based on a composite classifier which emphasizes certain patterns. More recently, Chang et al. [3] developed an approach to perform the articulatory labelling of a speech database through a connexionnist method. The method removed spectral vectors which were badly identified after a first learning phase. Thus, in a second learning phase, the strategy created multilayer perceptrons by only using the correctly identified spectral vectors.

2. Corpus material

The corpus 'Bref 120' for developing and evaluating speech recognition systems is used [4]. In this corpus, 120 adults read sentences taken from the newspaper 'Le Monde', and in total 66553 sentences are available. Recorded speech (sampling frequency 16kHz) and texts of the different sentences are available. By using a phonetic annotation procedure, the phonemes and their duration are defined and calculated for every sentence, excepted for the sentences containing proper nouns. Hence, 56519 sentences are annotated phonetically.

In this study, Bref 120 is shared in two different corpora for learning (A) and evaluation (B) stages. Both corpora are annotated phonetically with two different conditions. The first does not take into account the phonetic context and the second takes into account the context. Without context, the corpus is annotated with 36 labels. With context, the unvoiced stops (/p,t,k/) and fricatives (/f,s,ʃ/) are annotated as a function of the following vowels. In total, there are six different classes of context and the corpus is annotated with 66 labels. The classes are defined as a function of the following vowel features.

3. Elitist approach

As already mentioned, the elitist approach is based on a speech recognizer and a Dynamic Time Warping (DTW) algorithm. The speech recognizer is involved in two phases. The first, called 'training phase', creates the HMM-based speech sound models. The HMM models have three states, a simple left-right topology and a mixture of 64 gaussians. The grammar used is a simple phoneme loop. The second, called 'recognition phase', uses the speech recognizer and the HMM defined in the training phase to detect speech sounds in acoustic signals. To perform the training and the recognition phases, the software Espere developed by the Speech Group of Loria is used [5]. In

the sequel, we described the learning and evaluation stages of the elitist approach.

3.1. Learning stage

During the learning stage, the corpus A with both conditions of phonetic annotations is used. The goal of this stage is to train the HMM-based phoneme models. The strategy of the elitist approach is depicted in **figure 1**.

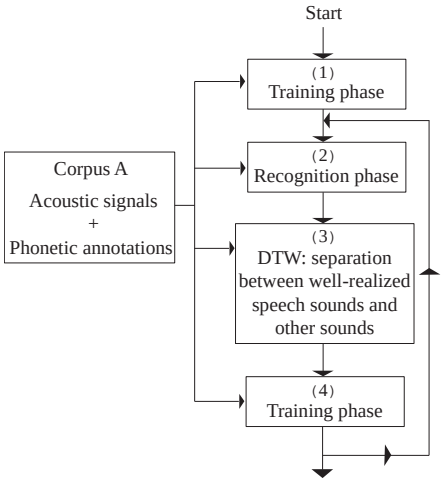

Figure 1: Scheme of the learing stage for the elitist approach.

(1) A training phase is performed to create the HMM-based phoneme models by using the acoustic signals and phonetic annotations of the corpus A.

(2) A speech recognition is carried out with the corpus A and the HMM models defined in stage (1).

(3) The DTW algorithm compares phonetic annotations of corpus A and the output of the speech recognition system. The wrongly-detected phonemes have their phonetic annotations changed. For instance, a phoneme annotated /p/ in the corpus A and wrongly-detected by the speech recognition system becomes /p'/. The phonetic annotation of the corpus A is changed after each iteration, and consequently the number of systematically well-detected speech sounds decreases.

(4) A training phase is again performed to create new HMM-based acoustic models of the systematically well-detected speech sounds and other speech sounds.

After this stage, an iterative procedure is carried out with the stages (2), (3) and (4) to improve the accuracy of the HMM for the systematically well-detected speech sounds and other speech sounds. At the first iteration, the HMM defined in stage (1) are used in stage (2), whereas at other iterations the HMM defined in stage (4) are used in stage (2).

To sum up, we design an elitist learning of HMM that makes very reliable sound models emerge. The learning is iterated by feeding sounds identified correctly at the previous iteration into the learning algorithm. Thus, the training phase produces models for the systematically well-detected speech sounds on the one hand, and standard models for the other sounds on the other hand.

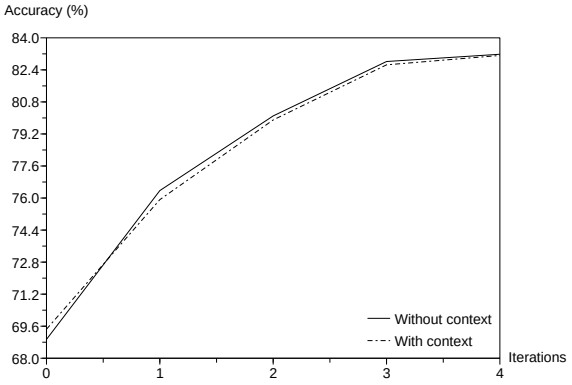

Figure 2: Learning stage: the accuracy (%) represents the rate of systematically well-detected speech sounds which are still correctly identified at iteration (n). The speech recognizer at iteration (n) uses the HMM-models of the systematically well-detected speech sounds trained with the sounds labeled as 'well-detected' at the iteration (n-1). The corpus A is used.

3.2. Evaluation stage

The performances of the elitist approach is evaluated with the corpus B. For this purpose, a speech recognition is carried out on the corpus B with HMM-based phoneme models of systematically well-detected speech sounds and models for other speech sounds. Afterwards, the DTW algorithm compares the output of the speech recognizer with the phonetic annotations of the corpus. Hence, the performance of the elitist approach can be estimated.

It is expected that the rate of trigger action for the HMM corresponding to the well-detected speech sounds is high while the rate of errors is small.

4. Results

4.1. Learning stage

During the learning stage, the elitist approach iterates on phonemes systematically well-detected. This means that the number of speech sounds systematically well-detected decreases at every iteration. After a few iterations, the speech sounds systematically well-detected are expected to have well marked acoustic cues.

The performance of the learning phase is evaluated by the identification rate (or accuracy) of previously well-detected speech sounds, and the percentage of systematically well-detected phonemes available in the corpus A. The higher the identification rate and the percentage of items, the better the performance of the learning phase.

At the output of the recognition phase (stage (2) of **figure 1**), the accuracy of the systematically well-detected speech sounds is measured at each iteration. Thanks to the DTW algorithm, the numbers of well-detected speech sounds (Ok), insertions (Ins), substitutions (Sub) and omissions (Omi) are known. Thus, the accuracy of the systematically well-detected speech sounds can be calculated by:

$$Accuracy = \frac{Ok - Ins}{Sub + Omi + Ok} \quad (1)$$

When the contextual models are used, the consonants

/p,t,k,f,s,ʃ/ are labeled with the name of the following vowels. However, for calculating the accuracy, if one element of a class is detected as another element of the same class, then this element is considered as well-detected (Ok) and not as substituted (Sub).

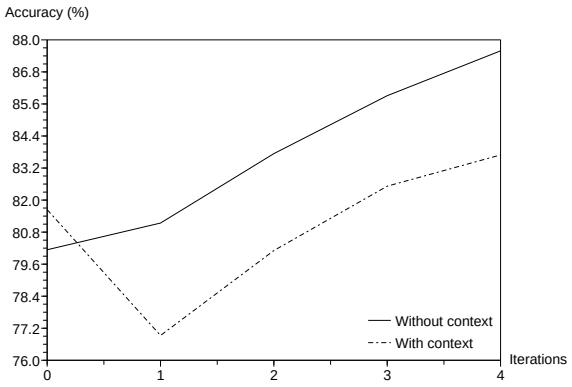

Figure 3: Learning stage: accuracy (%) of the elitist approach for the speech sounds /p,t,k,f,s,ʃ/ as a function of iterations. The corpus A is used.

Figure 2 and **3** show the accuracy at the output of the recognition phase for all the phonemes and the specific phonemes (/p,t,k,f,s,ʃ/), respectively.

As expected at the first iteration (0), the contextual models give a better accuracy than the non contextual models. However at the other iterations (1, 2, 3 and 4), the contextual models give the best accuracy.

A statistical analysis was carried out with all the phonemes to compare both phonetic annotation conditions (**figure 2**). The analysis estimated that there are significant differences between the contextual and non contextual models at the iteration 0, 1, 2 and 3. At the iteration 4, the statistical analysis estimated that they are no significant differences between both conditions of phonetic annotations.

With the specific phonemes (/p,t,k,f,s,ʃ/), **figure 3** shows that the accuracy increases at each iteration when the non contextual models are used. However with context, the rate decreases significantly between the first and the second iteration. This can be explained by the number of items which is available in the corpus for every phoneme.

Indeed, the performance of the speech recognizers may depend on the number of items which are available in the corpus to train HMM-based phoneme models. It seems that at the first iteration, the number of items is high enough to create accurate HMM for each phoneme. At the second iteration, the phonetic annotation is changed and the number of systematically well-detected speech sounds decreased significantly (**figure 4**). For a few phonemes, the number of items is around one hundred. The small number of items seems to decrease the quality of the HMM-based acoustic models using 64 gaussians, and then affects the performance of the speech recognizer.

Figure 4 shows the percentage of items annotated as well-detected after each iteration in the corpus A. The elitist approach is more selective with context than without context. For instance with specific phonemes /p,t,k,f,s,ʃ/ at the second iteration, 86.4% of the phonemes are still annotated as well-detected when the non contextual models are used. At the same iteration but with context, 65.4% of the phonemes are annotated as well-detected.

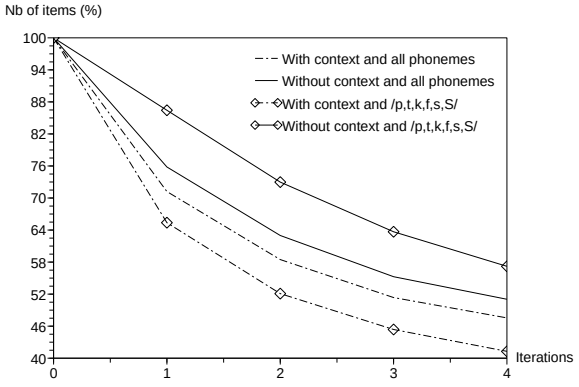

Figure 4: Percentage of items annotated as well-detected in the corpus A.

To sum up, the best identification rate (83%) of the systematically well-detected speech sounds for all phonemes is obtained at the fifth iteration (**figure 2**). This rate is the same for both phonetic annotation conditions. The identification rate is measured for the systematically well-detected speech sounds /p,t,k,f,s,ʃ/ (**figure 3**). The best accuracy (88%) is also obtained at the fifth iteration when the non contextual models are used. Finally, at the fifth iteration without context, 51% of all phonemes and 57% of the phonemes /p,t,k,f,s,ʃ/ are systematically well-detected (**figure 4**).

Hence, the non contextual HMM models created at the fifth iteration are used to perform the evaluation stage.

4.2. Evaluation stage

Speech recognition is performed with the HMM models created in the learning phase on well identified speech sounds and other speech sounds. The goal of the evaluation stage is to assess the rate of trigger action for the systematically well-detected speech sounds and the rate of false alarms.

Thus, we expect to have a significant identification rate of the systematically well-detected speech sounds with a minimum rate of errors. In the sequel, we are mainly interested in the identification rate of the specific speech sounds /p,t,k,f,s,ʃ/ in order to develop a speech enhancement technique [6]. We also assume that systematically well-detected speech sounds by speech recognition are well-realized sounds. Conversely, the other sounds are assumed to be not so well-realized sounds.

Since the speech recognizer uses the HMM model of the systematically well-detected and other speech sounds, the global rate of correct detection of the elitist approach is 79.2% on average over the unvoiced speech sounds (**table 1** and **2**). **Table 1** shows the percentage of unvoiced speech sounds identified as well-realized /p,t,k,f,s,ʃ/, and **table 2** shows the percentage of unvoiced speech sounds identified as not so well-realized /p',t',k',f',s',ʃ'/. Thus, the global rate of 79.2% is shared into a rate of 55.3% where the unvoiced speech sounds are identified as well-realized (**table 1**), and a rate of 23.9% where the unvoiced speech sounds are identified as not so well-realized (**table 2**).

The rate of false alarms is rather small for the unvoiced stops and fricatives considered as well-realized (**table 1**). For

		Detected models					
		/p/	/t/	/k/	/f/	/s/	/ʃ/
Speech sounds	/p/	44.88	1.23	0.49	0.32		
	/t/	1.63	55.79	0.53	0.13	0.44	
	/k/	0.29	0.88	47.98	0.13	0.39	
	/f/				69.71		0.27
	/s/				1.59	59.36	0.19
	/ʃ/				0.38	0.31	54.31
	/b/	0.64	0.19				
	/d/	0.18	1				
	/g/	0.12		1.43			
	/v/				1.08		
	/z/			0.23	0.24	1.06	
	/ʒ/						1.13
	/Others/	0.07	0.06	0.1	0.12	0.05	0.03
Insertions		1.62	1.15	2.93	5.57	1.06	2.77

Table 1: Rate (%) of trigger action for the HMM-based models of the well-realized unvoiced speech sounds. The corpus B is used. Only the percentages above 0.01% are mentioned.

instance, 1.63% of the speech sounds /t/ are identified as /p/ by the elitist approach. The rate of false alarms is 0.82% on average.

The rate of false alarms within the not so well-realized unvoiced speech sounds is rather small and is about 0.7% (**table 2**).

		Detected models					
		/p'/	/t'/	/k'/	/f'/	/s'/	/ʃ'/
Speech sounds	/p/	34.8	1.71	0.93	0.24	0.27	0.06
	/t/	3.66	18.86	2.99	0.29	0.83	0.24
	/k/	1.26	1.44	26.88	0.27	0.51	0.21
	/f/	0.28	0.23	0.27	17	0.7	0.41
	/s/	0.12	0.24	0.37	1.04	21.79	0.82
	/ʃ/	0.04	0.2	0.29	0.58	1.61	24.3

Table 2: Rate (%) of trigger action for the HMM-based models of the not so well-realized unvoiced speech sounds. The corpus B is used. Only the percentages above 0.01% are mentioned.

To sum up, the speech recognizer identifies 55% of the speech sounds as well-realized, with a small rate of errors. Hence, the elitist approach can extract automatically speech sounds, assumed to be well-realized, with high confidence. With the aim of developing a speech enhancement technique, this is an important result. Indeed, we will be able to perform modifications with high confidence on the speech sounds detected by the speech recognizer using the HMM models of the systematically well-detected speech sounds of the learning phase.

5. Perspectives

We have assumed that the systematically well-detected speech sounds have well marked phonetic features. This is why we have created HMM-based acoustic models for the supposed well-realized speech sounds. Our next work is to prove that the well-detected speech sounds have well marked phonetic features. To perform this work, a system devoted for the identification of stop consonants will be used [1].

This system exploits acoustic detectors designed in the study of Bonneau et al. to evaluate acoustic cues provided by burst and formant transitions. Since spectral characteristics of burst release are more relevant to identify stop consonants, formants are tracked automatically and a segmentation algorithm separates the burst release from the frication noise. Spectral cues, particularly the emergence and the frequency of the most prominent peak, turned out to be very efficient to detect strong cues.

The work will consist in studying the spectral cues of the well-realized speech sounds.

6. Conclusions

Promising results were obtained with the elitist approach for extracting automatically unvoiced stops and fricatives. On average, 55% of the specific speech sounds are classified as well-realized speech sounds with a small percentage of errors.

A next work will consist in studying the acoustic cues of the speech sounds assumed as well-realized and not so well-realized speech sounds. Finally, perceptual tests will be carried out for testing if the amplification of the well-detected unvoiced stops and fricatives can improve speech intelligibility.

7. Acknowledgement

Partly funded by Voice Web Initiative - HP Philantropy and Education in Europe.

8. References

[1] Bonneau, A., Coste-Marquis, S., and Laprie, Y. “Strong cues for identifying well-realized phonetic features”, Intern. Congr. of Phon. Sci., 1995, Stockholm, Sweden, p 144-147.

[2] Schwenk, H. “Using boosting to improve a hybrid HMM/neural network speech recognizer”, ICASSP, 1999, Arizona, United States, p 1009-1012.

[3] Chang, S., Greenberg, S., and Wester, M. “An elitist approach to articulatory-acoustic feature classification”, Eurospeech, 2001, CD.

[4] Lamel, L.F., Gauvain, J.L., and Esknazi, M. “BREF, a Large Vocabulary Spoken Corpus for French”, EUROSPEECH, 1991, Genoa, Italia, p 505-508.

[5] Fohr, D., Mella, O., and Antoine, C. “The automatic speech recognition engine ESPERE : experiments on telephone speech”, ICSLP, 2000, Pkin, China.

[6] Colotte, V. “Techniques d’analyse et de synthse de la parole appliquee a l’apprentissage des langues”, PhD dissertation, 2002, Loria, Nancy, France.