AUTOMATIC PHONETIC SEGMENTATION AND LABELLING OF SPONTANEOUS SPEECH

Luis Pinto Coelho[1] *y Daniela Braga*[2]

[1] Instituto Politécnico do Porto, [2] Universidade da Coruña

ABSTRACT

In this paper a tool for automatic segmentation and labeling of spontaneous speech is presented. It is developed and specially tuned for the European Portuguese (EP) language but simple changes are needed to convert it to other languages. The main purpose of this system is to quickly produce a high quality output of phonetic labels and related time boundaries using as input the speech signal only. Our motivation was the development of a tool that could help to create new voices for TTS systems without and special previous text selection concerns. A quick talk or a quick reading of a randomly selected text are the targets for the use of our system. The evaluation of the presented system gave up to 88.3% of accuracy in phonetic alignment considering a 20ms temporal error.

1. INTRODUCTION

The quick development of speech systems created a demand for new and better speech databases (using new voices, new dialects, new special features to consider, etc.), often with phonetic level annotation information (and others). This trend re-enforces the importance of automatic segmentation and annotation tools because of the drastic time and cost reduction in the development of speech corpora even when some little human action is needed. For the Portuguese language, in spite of its world wide usage, the number of available resources is still low and unadjusted, justifying this way an effort on the improvement and development of databases and tools. An accurate method of automatic phonetic transcription can thus facilitate the development of TTS and ASR systems for novel material, both within and across languages, as well as increase robustness with respect to acoustic interference and variation in speaking style and pronunciation.

In section 2 an overview of the system is given including a brief description of the used corpus.

In section 3 the final results and some conclusions are presented (as a whole due to available space and esthetical considerations).

2. SYSTEM DEVELOPMENT

The system's global structure is shown in figure 1. The annotation tool uses as input a speech signal that will be decoded, on a first stage, by an HMM based recognition engine (other authors try to test several phonetic transcriptions obtained from orthography). The resulting phoneme sequence and the given audio will then be processed by a segmentation engine, also HMM based. Both engines have independent parameters but share a common phoneme inventory and a set of language and acoustic models previously trained. The use of an HMM framework gives the possibility of obtaining in a single stage, both phonetic sequence and temporal boundaries however, as will be shown, the demands of each task are different and require independent adjustments to effectively produce the desired results. Hence the use of two stages is mandatory.

The speech signal, labels and related boundaries are finally analysed as a whole in a refinement stage for producing high quality results. The duration of the phones, the acoustical/frequential positioning of time boundaries, the voiced/unvoiced behaviour, the pitch periods, among others aspects are considered in this final stage. After refinement the system outputs a list of phones with well defined temporal marks for begin and end.

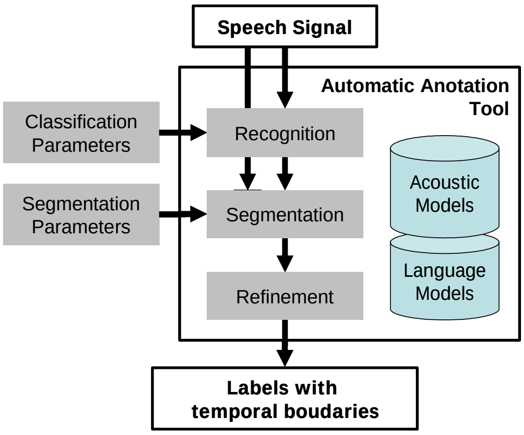

Figure 1. System overview.

2.1. Used Corpora

For system training, the FEUP/IPB speech corpus was used on a first step. This is a high quality database, recorded using professional equipment in a professional studio. It has nearly 100 minutes of speech from a single professional speaker and has phonetic level annotation [4].

On a second step, to increase robustness, the ProGmatica corpus [5] has been used. ProGmatica consists of broadcasted television materials, such as interviews, talk shows and political debates, in which natural spontaneous speech occurs. It was developed to study speech acts, using the well known John Searle's topology, and it is still an ongoing work. Around 20 minutes of phonetically hand labelled speech were selected considering the largest variety of voices and the inexistence of speech/voice overlaps.

The ProGmatica corpus has also been used for system evaluation. For this another 20 minutes were chosen using the same criteria.

2.2. HMM Framework

This system is based on a HMM framework with 50 models (20 consonants, 18 vowels, 8 diphthongs, 1 oclusion for voiced plosives, 1 occlusion for unvoiced plosives, 1 pause, 1 aspiration) with the same basic left to right topology with jumps (by quotient 9.7% better than the no-jumps topology) and 3 states in most cases. The first 38 models constitute the phonetic base for standard EP.

The errors found in vowel separation in diphthongs were significantly greater than those encountered in other situations. The soft formant frequencies transition make it hard, to humans and machines, to accurately define the inner frontier. These errors affect the overall performance for each independent phoneme occurrence and decrease significantly the global system quality because vowels represent about 50% of phoneme occurrences in EP. As EP has a large variety and high frequency of diphthongs, when compared to other languages (English for example), a study has been made in order to find the performance difference that would occur if these phonetic units were considered independently or as a whole. Adding 8 extra HMM with 5 states (9.3% improvement relative to 3 states), representing the EP main diphthongs, resulted on an increase of 3.4% in boundaries localization comparing to the independent phoneme (vowels) results and no benefits were noticed in classification.

For plosives, which represent nearly 20% of the average phonetic occurrences in EP, distinct models were created for voiced and unvoiced cases. This can give an increase of 1.1% in segmentation results and 0.3% in classification.

For non phonetic events two extra models are used to represent aspiration and pause.

2.3. Trainning

Since this approach is based on models that describe a speech signal in time, it is possible to simultaneously perform both segmentation and labeling at the same time. The pursue for an optimal configuration for the annotation system showed, by the analysis of the results, that the two implied tasks demand incompatible parameters. The variation of a parameter can improve performance on one hand (better classification) and decrease it on another hand (worst segmentation). Thus, two function oriented sets of configuration parameters were developed, one for segmentation and another for labeling.

Varying the analysis window length and the overlap time parameters several results can be obtained. Table 1 resumes this analysis showing the correctness for recognition and error rate for segmentation for a system with 12MFCC. Insertions, deletions and substitution errors (these last only considered for classifying) are included.

Overlap	Window length (ms)		
	15	20	25
2.5	**76.5** / 65.2	76.3 / 63.2	75.3 / 61.0
5.0	75.0 / **69.7**	74.7 / 67.2	73.8 / 65.7
7.5	72.3 / 64.9	72.5 / 64.5	70.0 / 63.9
10.0	68.4 / 62.9	69.3 / 62.7	68.6 / 61.1

Table 1. Recognition/Segmentation results in percentage for several windows and overlap times.

States	Class. Acc. (%)	Segment. Acc. (%)
3	0,00% baseline	0,00% baseline
4	7,01%	-9,03%
5	2,75%	-7,77%
6	0,00%	-12,46%
7	0,00%	-12,43%

Table 2: Results varying the model's number of states

Mix.	Class. Acc. (%)	Segment. Acc. (%)
3	0,00% baseline	0,00% baseline
4	15,18%	4,96%
5	19,41%	7,67%
6	22,92%	12,89%
7	23,85%	11,15%

Table 3: Results varying the state's number of mixtures

Feature	Class. Acc. (%)	Segment. Acc. (%)
MFCC	0,00% baseline	0,00% baseline
MFCC_E	31,39%	18,28%
MFCC_0	28,93%	13,79%
MFCC_ED	57,55%	39,05%
MFCC_DA	64,98%	45,61%
MFCC_EDA	70,88%	50,81%

Table 4: Results varying the feature vectors (E for Energy, D for Delta, A for Acceleration)

Tables 2, 3 and 4 show the results for others parameters variations in relation to a very simple set of initial parameters that served as a baseline (25ms window, 10ms overlap, 12MFCCs).

Finally, considering all the obtained results, two sets of parameters were defined. For segmentation, 15 ms frames with 5 ms overlap have been used. The feature vector was composed by 16 MFCCs plus Energy, delta and acceleration coefficients, for each state 7 Gaussian mixtures. For classifying, 25 ms frames with 7.5 ms overlap, 14 MFCCs plus Energy, delta coefficients and 5 Gaussian components per state have been used.

To reduce the complexity in decoding the phonetic sequence a language model was also developed. This first language model, also used in the previously described analysis, was based on simple rules that benefict from the characteristic consonant-vowel (CV) alternance of EP. The sequences CCV and VV are also common. This model was then extended to a bigram in order to include exceptions. With this, the results had an accuracy of 78.67% for classification, 77.97% for a 10 ms tolerance zone and 84.33% for a 20ms tolerance zone.

In the development of these systems, when the final accuracy values were similar for two different parameter sets, the one with more insertion errors was preferred. This kind of error is easier to be corrected. So, with this in mind, the above accuracy values have almost no deletion errors.

2.3. Training

The set of HMM was first initialized using the FEUP/IPB corpus in order to obtain a solid phoneme description. After, the speaker dependent highly constrained HMM set need to be re-trained and adapted to other voices and to non ideal studio conditions. This way, the ProGmatica corpus was used.

2.4. Refinement

In the last stage, the annotation tool analyses the results so far and tries to improve boundaries' positions and classification errors.

The given boundaries are analyzed using cues given by acoustical features extracted from the speech signal. For the selection of the most interesting feature that could give the best representation of the acoustical information, a study has been made giving a special emphasis on autoregressive moving average models (ARMA). These models are based on the generic equation 1, representing the denominator an AR model (with $b_l=0$) and the numerator a MA model ($a_k=0$).

$$H(z) = G\frac{1+\sum_{l=1}^{q} b_l z^{-1}}{1+\sum_{k=1}^{p} a_k z^{-1}} \qquad (1)$$

The number of AR and MA coefficients considered to effectively describe the speech signal has been independently analyzed. The results in table 5 have been obtained using 16KHz speech signals, 20ms windows with Hamming mask and 10ms overlapping.

Phon.	MA	AR	Avg. Error	Max. Error
m, n, J	16	4	0.53%	3.07%
	4	16	1.51%	12.13%
	16	16	0.19%	18.60%
a, p, s	16	4	2,37%	16.54%
	4	16	1,02%	9.72%
	16	16	0.18%	6.85%

Table 5. Prediction Error by Model Type

The all-pole models, commonly used, can give a very good representation of most phonemes but, when applied to nasals and some unvoiced sounds, the performance decreases a lot. The use of some MA coefficients increases the results and comes to solve this problem. The benefice of adding new poles or zeros can only be observed when the description capacity of the model is not reached. After this, some slight decrease can be noticed.

Using the described models, speech feature vectors have been composed using 4 zeros from de MA model, 16 poles from the AR model, signal energy and the dynamic related values delta and acceleration. The signal is analyzed from 20ms before the given boundary to 20ms after using 10ms segments with a Hamming window and a 2.5ms overlap. The Euclidean distance is calculated for every two successive vectors and the likelihood is observed. The phonetic boundary confidence is based on a threshold level which is calculated locally. The standard deviation is also used to create a threshold band. Instead of having only a simple frontier line that limits two phonemes in a multi-dimensional space, two limits are used, one for a raising distance and another for a decreasing distance (hysteresis classifier). This behavior has been implemented in order to reduce the number of false identified boundaries.

The final phonetic sequence is yet reviewed. The too big and too small occurrences (very rarely) are deleted and the duration of each phoneme is tested against pre-known EP knowledge. In this case the error in boundaries is calculated for each phone and compared with its reference duration. An error of 15ms in a /d/, which has an average duration of 15.7ms, would be much worst than the same error in an /a/, which has an average duration of 112.3ms. Table 6 shows the average errors and the average durations.

The log probability associated with each phoneme is also used in order to remove some errors. A problem comes up in the deletions of phonemes. A phone has two boundaries and, assuming that the phonetic sequence is composed by contiguous segments, when a phone is deleted, one of the frontiers must be moved. This can be made by readjusting the frontier of the previous phone or by repositioning the beginning of the next phone. The use of rigid rules for the two operations has been examined. The assimilation of the phone by the previous phone was more interesting but the differences between the two did not dictate a rule that could be generalized.

Symb .	Avg. Dur. (ms)	Avg. Err. (%)	Symb .	Avg. Dur. (ms)	Avg. Err. (%)
a	112,3	0,96%	ocl	51,6	0,29%
asp	292,0	1,01%	O	94,3	2,77%
@	51,8	1,35%	p	24,3	7,96%
d	15,7	2,32%	r	35,8	0,45%
E	77,2	1,66%	R	79,8	3,40%
E	83,5	1,02%	s	99,7	3,30%
i	63,0	0,65%	S	83,1	0,45%
i_	85,3	3,96%	6	63,4	1,62%
j	32,8	5,12%	t	31,5	0,95%
k	35,5	5,49%	u	48,6	0,21%
l	49,1	1,19%	u_	77,9	2,86%
L	81,7	2,71%	v	57,4	0,70%
m	58,9	0,15%	w	25,5	1,59%
o	82,8	2,13%	Z	82,7	0,27%

Table 6. Average duration for reference phonemes and average error in segmentation.

For improving this task an algorithm has been created. For deleting phones, the duration of the two neighbor phones are analyzed. The phone to change will be the one that, after deletion, can give a set of durations closer to those in the knowledge database. The described operations can be seen on figure 2 in which are represented the two deletions' possibilities.

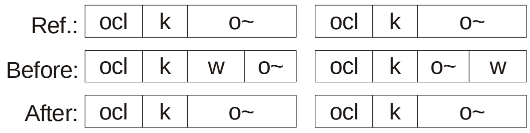

Figure 2. Phone deletion at left and right.

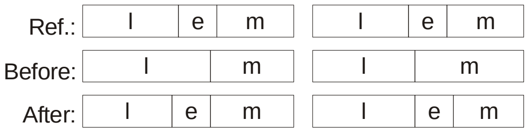

Figure 3. Phone insertion at left and right.

For insertions of phones in the phonetic sequence (corrections of deletion errors) a similar problem occurs. In this way, similar rules are used. The new label that will be inserted is obtained by analyzing context and a grammar. The durations of the phones that will be separated by the new phone are analyzed and compared with the average occurrences in the database. The frontier which is changed is the one that can give a better duration similarity with the know phones' durations. The duration of the new phone is based on the average duration of the same phones in the database and it is adjusted by a factor resulting from the comparison of the durations of the neighbor phones. This operation is shown in figure 3.

3. RESULTS AND CONCLUSION

With the described process, the obtained results were, for segmentation accuracy, 65.3% for a 10ms interval, 79.2% for a 15ms interval and 88.3% for a 20ms interval as can be seen in figure 4.

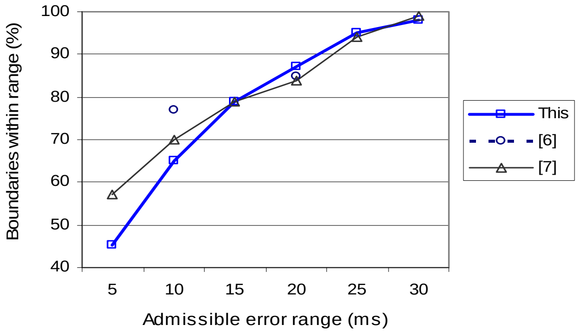

Figure 4. Final results and comparison

In this paper, an automatic annotation system has been presented. Regarding the use of natural speech recorded in real situations and trying to ease the quick creation of speech databases the system take as entry information only the speech signal. Some refinement procedures based on signal analysis can give an extra refinement to the boundaries localization. These results are close to or better than those reported by other authors that also worked with EP but not with spontaneous speech [6][7].

4. REFERENCES

[1] A. Sethy, S. Narayanan, "Refined Speech Segmentation for Concatenative Speech Synthesis", Proc.of ICSLP, Denver, 2002.

[2] A. Ljolje, M. Riley, "Automatic Segmentation and Labelling of Speech", Proc. of ICASSP, p473-476, 1991.

[3] J. Hosom, "Automatic Time Alignment of Phonemes using Acoustic-Phonetic Information", PhD Thesis, OGI, 2000

[4] J. Teixeira, D. Freitas, D. Braga, M. J. Barros, V. Latsch, "Phonetic Events from the Labeling of the European Portuguese Database for Speech Synthesis, FEUP/IPB-DB", Proc. of EuroSpeech'01, Denmark, 2001

[5] D. Braga, J. P. Teixeira, L. Coelho, D. Freitas, "ProGmatica: Uma base de dados prosódica pragmaticamente orientada em Português Europeu", Actas do XXI Encontro da Associação Portuguesa de Linguística, Lisboa, 2006

[6] P. Carvalho, "Determinacão Automática de Segmentos para Síntese de Fala por Concatenação", PhD thesis, Universidade Técnica de Lisboa - IST, 2004.

[7] L. Coelho, "Etiquetagem Automática de Sinais de Fala: Classificação e Anotação", Tese de mestrado, FEUP, 2005