Automatic identification of French regional accent

Maëlys Salingre

IRIT/RR--2017--13--FR

Intership under the direction Jérôme Farinas (IRIT) and Stéphane Rabant (Authôt)

Work completed in SAMOVA team of Institut de Recherche en Informatique de Toulouse and Authôt society at Ivry-Sur-Seine.

Automatic identification of French regional accent

Maëlys Salingre, Jérôme Farinas, Stéphane Rabant

This study presents an implementation of an automatic tool to identify the regional dialects of French. GMM-UBM modeling has been developed on the KALDI platform. We arrive at 40\% of regional accent identification on 31 survey points in France, Switzerland and Belgium, using acoustic parameters. The addition of the fundamental frequency allows a slight gain on the results. But the contribution of rhythm and intonation separate models did not improve performances, at most it allows for more coherent regional grouping.

Keywords: automatic language recognition, intonation modeling, rhythm modeling, French regional accent, prosody

Introduction

The performance of an Automatic Speech Recognition (ASR) system is sensible to speaker variation: age, sex, origin, possible language pathologies, etc. Among these variation factors, the origin of a speaker can be linguistically characterized through their regional accent or dialect. Being able to identify a speaker's regional accent and using a specifically trained model can reduce the WER up to 20% (Humphries & Woodland 1997).

Dialect and accent recognition methods are largely inspired by language recognition and speaker recognition methods. PRLM (Phone Recognizer Language Model) was one of the first method used to tackle the problem of accent and dialect recognition (Zissman et al. 1996). The ACCDIST distance matrix (Barry et al. 1989) has also been widely used especially for English vowels (Ferragne 2008). More recently GMM-UBM and ivectors who showed great results for speaker recognition became popular for dialect recognition (Lazaridis et al. 2014, Bahari et al. 2014). All these methods have in common that they rely only on acoustic features.

In this study, we will see how pitch and prosodic features can improve performances for French regional accent detection compared to just acoustic features.

1. Previous studies

The creation of large corpora of spoken French has made possible researches in French regional accent recognition in recent years.

Boula de Mareüil et al. (2008) studied human and computer recognition of foreign and regional accents in French. They showed that native French speakers could recognize three main accents out of 12: Northern French (standard French), Southern French and Swiss French. They concluded

using a decision tree that F2 values for the phoneme /ɔ/, schwa elision and post-nasalization were the most informative for identifying regional accents.

Lazaridis et al. (2014) used GMM-UBM with EM (Expectation Maximization) and MAP (Maximum A Posteriori) as well as TV (Total Variability) and ivectors to identify four regional accents of Swiss French. They obtained a best accuracy of 38% using TV outperforming the GMM baseline by 5 points.

2. Presentation of the corpus

We used the PFC (*Phonologie du Français Contemporain*) database for the experiment (Durand et al. 2002, 2009, available online at http://www.projet-pfc.net/). It is made up of several corpora that were recorded in specific regions. For each investigation point between 6 and 15 speakers were recorded. The recordings include a word list and a text reading as well as free and guided speech.

The corpus was restricted to 31 investigation points in Metropolitan France, Switzerland and Belgium:

Table 1: Investigation points

Point	Abbreviation	Nb of speakers	Point	Abbreviation	Nb of Speakers
Aix-Marseille	aix	8	Marseille-Centre	mar	10
Bar-sur-Aube	bar	10	Montreuil	mon	8
Béarn	bea	7	Nantes	nan	10
Biarritz	bia	12	Neuchâtel	neu	13
Brécey	bre	11	Nice	nic	8
Brunoy	bru	10	Nyon	nyo	12
Cussac	cus	15	Ogéviller	oge	11
Dijon	dij	8	Paris-Centre	par	12
Domfrontais	dom	12	Puteaux-Courbevoie	put	6
Douzens	dou	10	Roanne	roa	8
Gembloux	gem	12	Rodez	rod	8
Genève	gen	9	Salles-Curan	sal	12
Grenoble	gre	9	Toulouse	tol	14
Lacaune	lac	13	Tournai	tou	12
Liège	lie	12	Vendée	ven	8
Lyon	lyo	10			

Here is a map showing where the investigation points are localized:

Figure 1: Map with the 31 investigation points used (source: Google maps)

Only recordings of text reading were used to reduce variability due to the vocabulary. Recordings with too much background noise or where speakers had strong reading difficulties were not used. There was a bit more than 14 hours of speech in total.

3. Experiment

The experiment was conducted using the Kaldi ASR toolkit (Povey at al. 2011) LRE07 recipe. As Lazaridis et al. (2014), it uses GMM-UBM and ivectors. Using 128 Gaussians for the GMM and 600 dimensions ivectors gave the best results.

One speaker was chosen randomly from each investigation point to use as test and so there was around 13 hours of train speech and 2h15 of test speech.

All sound files were sampled down to 8,000Hz and converted to mono. The train files were cut into smaller files of a maximum length of 30 seconds. The test files were semi-automatically cut into 3, 10 and 30 second long files by using the available TextGrid annotation files to determine pauses.

3.1. Acoustic features

We started with the default recipe where only MFCC are used. Vocal tract length normalization (VTLN) and cepstral mean value normalization (CMVN) are applied to the MFCC and the deltas extracted to train the GMM.

Table 2: Results for acoustic features

	3 seconds	10 seconds	30 seconds	Total
Pmiss	0.7476	0.6241	0.5061	0.6993
Pfa	0.0249	0.0208	0.0175	0.0233
Cost	0.3863	0.3225	0.2618	0.3613

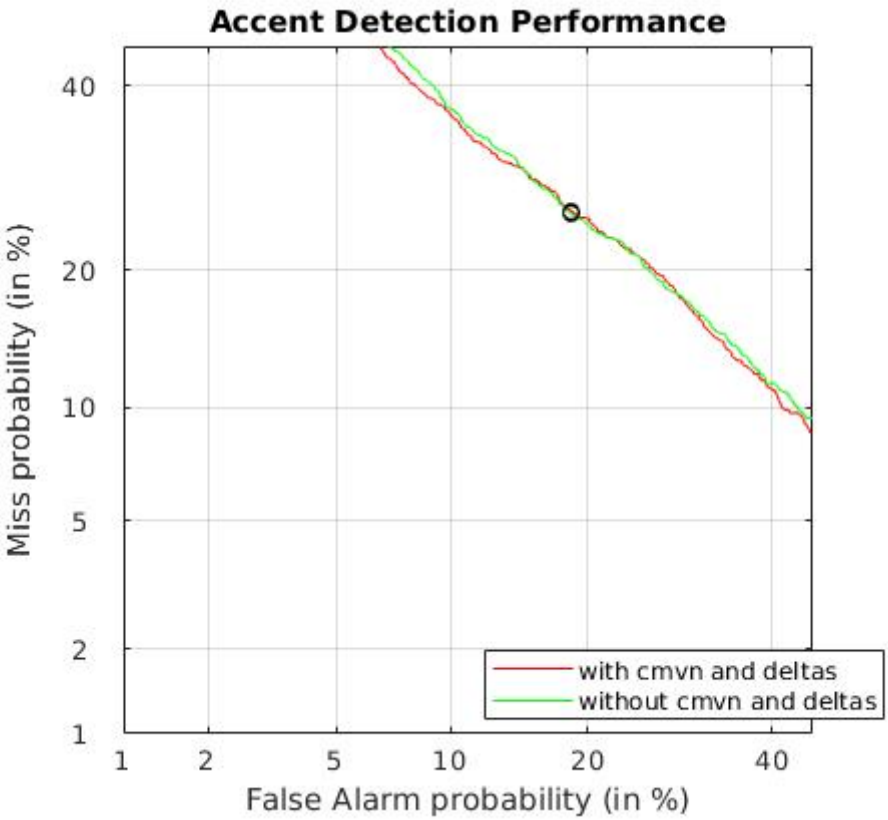
We also conducted the experiment without CMVN and the deltas for practical reasons: it is easier to add prosodic features to the Kaldi recipe without them.

Table 3: Results for acoustic features without CMVN and deltas

	3 seconds	10 seconds	30 seconds	Total
Pmiss	0.7313	0.6050	0.5850	0.6847
Pfa	0.0244	0.0202	0.0202	0.0228
Cost	0.3778	0.3126	0.3026	0.3538

Here is the DET (Detection Error Trade-off) curve for both runs. The circles represent the minimum cost points for each curve.

Figure 2: DET curve plot for acoustic features



The performance with CMVN and deltas is slightly better than without however there is not much difference. And so we believe that not using CMVN and deltas will not impact the results much.

3.2. Pitch features

Using Kaldi, the following pitch features were added to the MFCC: warped NCCF (normalized cross correlation function), log-pitch with POV (probability of voicing) -weighted mean subtraction over 1.5 second window and delta feature computed on raw log pitch.

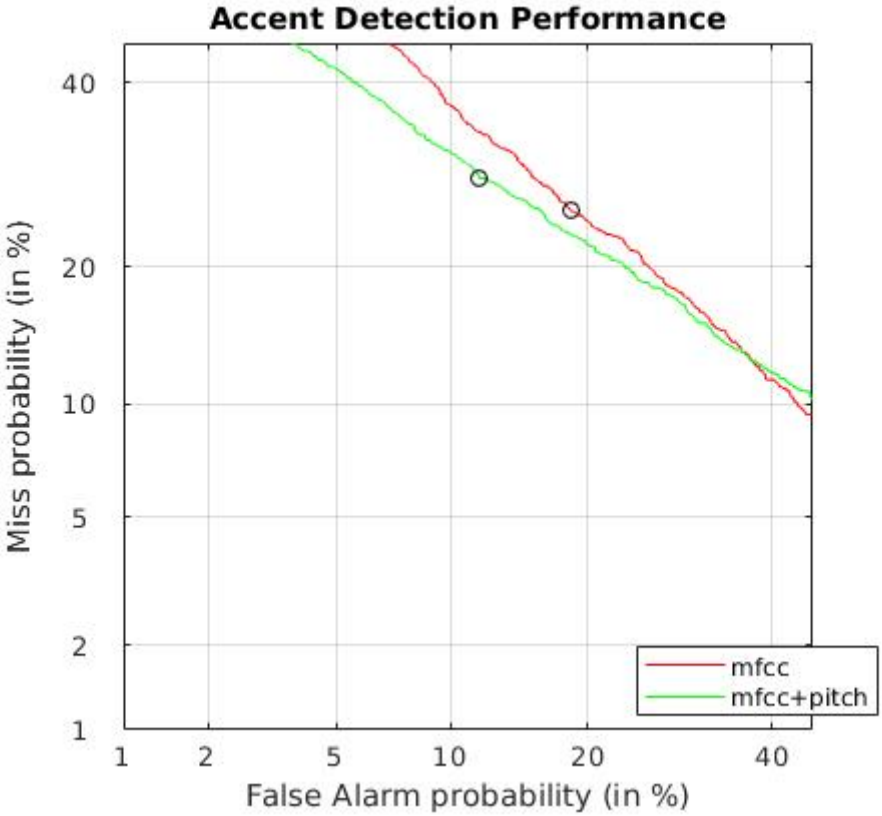
Here are the results:

Table 4: Results for acoustic features and pitch features

	3 seconds	10 seconds	30 seconds	Total
Pmiss	0.6281	0.5313	0.4889	0.5940
Pfa	0.0209	0.0177	0.0169	0.0198
Cost	0.3245	0.2745	0.2529	0.3069

Adding pitch features to the acoustic features reduces the miss probability up to 10 points. The gain in performance is especially noticeable for short test files.

Figure 3: DET curve plot for acoustic and pitch features



The DET curve plot confirms that using pitch features in addition to acoustic features improves accent detection performance.

3.3. Prosodic features

Intonation and rhythm features were extracted to create a prosody model. Utterances were automatically segmented into syllables and vowel nuclei with Praat using the Prosogram plugin (Mertens 2004). Smaller units, hereafter called acoustic events, were determined using the Forward Backward Divergence algorithm (André-Obrecht 1988). The main advantage of this segmentation method is that it does not require manual validation.

3.3.1. Intonation model

Pitch was extracted using the Kaldi pitch tracker (Ghahremani et al. 2014). Two sets of features were extracted. The first set was inspired by the RFC (rise/fall/connection) model (Taylor 1992). Here are the extracted features:

Table 5: RFC intonation features

Feature	Description
Rise amplitude (rise_amp)	Difference in Hz between the pitch peak and the pitch value at the beginning of the syllable
Normalized rise amplitude (rise_amp_norm)	Rise amplitude normalized with the mean pitch
Fall amplitude (fall_amp)	Difference in Hz between the pitch value at the end of the syllable and the pitch peak
Normalized fall amplitude (fall_amp_norm)	Fall amplitude normalized with the mean pitch
Total amplitude (total_amp)	Difference in Hz between the pitch peak and the lowest pitch value in the syllable
Normalized total amplitude (total_amp_norm)	Total amplitude normalized with the mean pitch
Peak height (peak_height)	Value in Hz of the pitch peak
Normalized peak height (peak_height_norm)	Peak height normalized with the mean pitch
Position (pos)	Difference in seconds between the pitch peak and the beginning of the vowel nucleus
Normalized position (pos_norm)	Position normalized with the vowel nucleus duration
Rise duration (rise_dur)	Difference in seconds between the pitch peak and the beginning of the syllable
Normalized rise duration (rise_dur_norm)	Rise duration normalized with the syllable duration
Fall duration (fall_dur)	Difference in seconds between the end of the syllable and the pitch peak
Normalized fall duration (fall_dur_norm)	Fall duration normalized with the syllable duration

For the second set, statistical features from Farinas (2002) and Rouas (2005) were used. The mean, variation, kurtosis and skewness were calculated for the pitch values over the syllable. The “maximum of accentuation” in Farinas (2002) and Rouas (2005) corresponds to the position in the RFC set.

A one-way ANOVA was conducted using R using each syllables intonation features as the dependent variables and the investigation point as the independent variable. All features were found to be significant ($p=0.00992$ for pos_norm and $p<2e^{-16}$ for all the other features). Only normalized features were selected for the experiment. The RFC set was comprised of total_amp_norm, rise_dur_norm and peak_height_norm while the statistical set of pos_norm, kurt and skew. Since the syllable segmentation was automatically done, some syllables were only one 10ms frame long and so variation was not calculable.

3.3.2. Rhythm model

The rhythm model was inspired by Farinas (2002). Hereafter we consider the parts of the syllable outside the vowel nucleus as consonants. The following features were extracted:

Table 6: Rhythm features

Feature	Description
Duration (dur)	Duration of the syllable in seconds
Vowel duration (dur_v)	Duration of the vowel nucleus in seconds
Normalized vowel duration (dur_v_norm)	Vowel duration normalized with the syllable duration
Consonant duration (dur_c)	Duration of all the consonants of the syllable in seconds
Normalized consonant duration (dur_c_norm)	Consonant duration normalized with the syllable duration
Complexity (comp)	Number of acoustic events in the syllable
Normalized complexity (comp)	Complexity normalized with the syllable duration
CV ratio in events (ratio_cv_events)	Number of consonantic events divided by the number of vocalic events
CV ratio in duration (ratio_cv_dur)	Consonant duration divided by vowel duration
Normalized CV ratio in events (ratio_cv_events_norm)	CV ration in events normalized with the syllable duration
Consonant complexity (comp_c)	Number of consonantic events
Normalized consonant complexity (comp_c_norm)	Consonant complexity normalized with the syllable duration
Consonantic events mean duration (dur_events_c)	Mean duration of consonantic events
Vocalic events mean duration (dur_events_v)	Mean duration of vocalic events

As for intonation features, a one-way ANOVA was conducted with R. All features were found to be significant ($p<2e^{-16}$) except for ratio_cv_dur ($p=0.174$). The selected features were dur_v_norm, ratio_cv_events and comp_c_norm.

3.3.3. Results

We conducted five experiments by adding different prosodic features to the acoustic and pitch features. The first three experiments used respectively the rhythm features, the statistical intonation features and the RFC intonation features. The last two experiments used rhythm and statistical intonation features and rhythm and RFC intonation features.

Table 7: Results for rhythm features

	3 seconds	10 seconds	30 seconds	Total
Pmiss	0.6967	0.6044	0.51	0.6623
Pfa	0.0232	0.0201	0.0176	0.0221
Cost	0.36	0.3123	0.2638	0.3422

Table 8: Results for statistical intonation features

	3 seconds	10 seconds	30 seconds	Total
Pmiss	0.8424	0.7468	0.6894	0.8068
Pfa	0.0281	0.0249	0.0238	0.0269
Cost	0.4352	0.3858	0.3566	0.4168

Table 9: Results for RFC intonation features

	3 seconds	10 seconds	30 seconds	Total
Pmiss	0.7275	0.6272	0.5756	0.6881
Pfa	0.0242	0.0209	0.0198	0.0229
Cost	0.3759	0.3241	0.2977	0.3555

Table 10: Results for rhythm and statistical intonation features

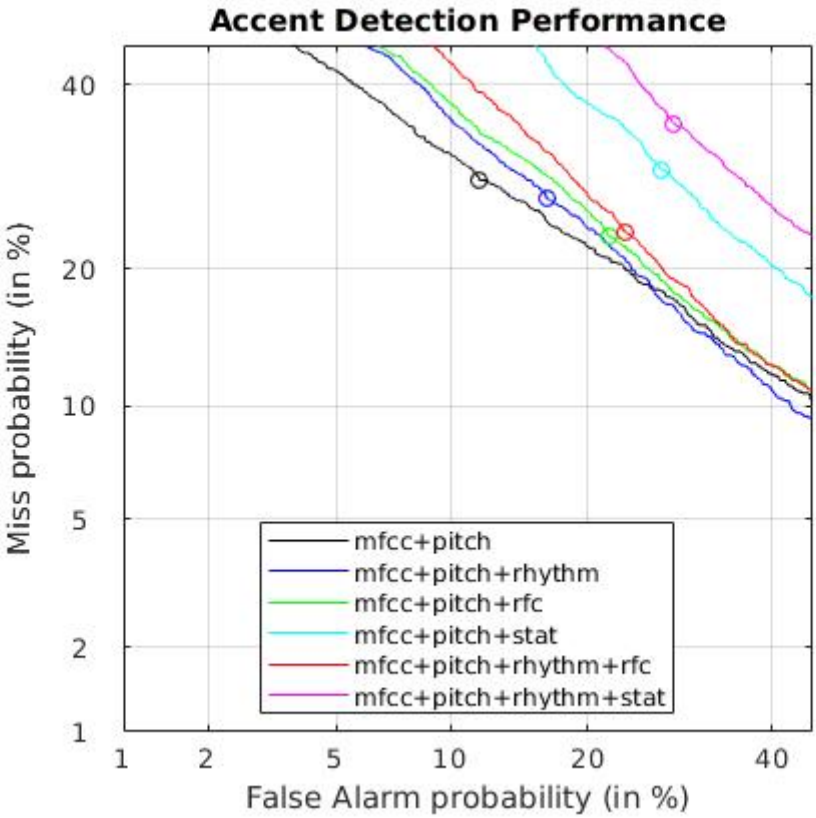
	3 seconds	10 seconds	30 seconds	Total
Pmiss	0.8676	0.8117	0.7656	0.8430
Pfa	0.0289	0.0271	0.0262	0.0281
Cost	0.4483	0.4194	0.3959	0.4356

Table 11: Results for rhythm and RFC intonation features

	3 seconds	10 seconds	30 seconds	Total
Pmiss	0.7585	0.6887	0.5961	0.7301
Pfa	0.0253	0.0230	0.0206	0.0243
Cost	0.3919	0.3558	0.3083	0.3772

Adding prosodic features does not improve performance comparing to acoustic and pitch features. Statistical intonation features especially degrade performances: the results are worse than with only acoustic features. Another interesting observation is that although rhythm features gave better results than RFC intonation features, using both set of features at the same time degraded accent detection performance more than with just RFC features.

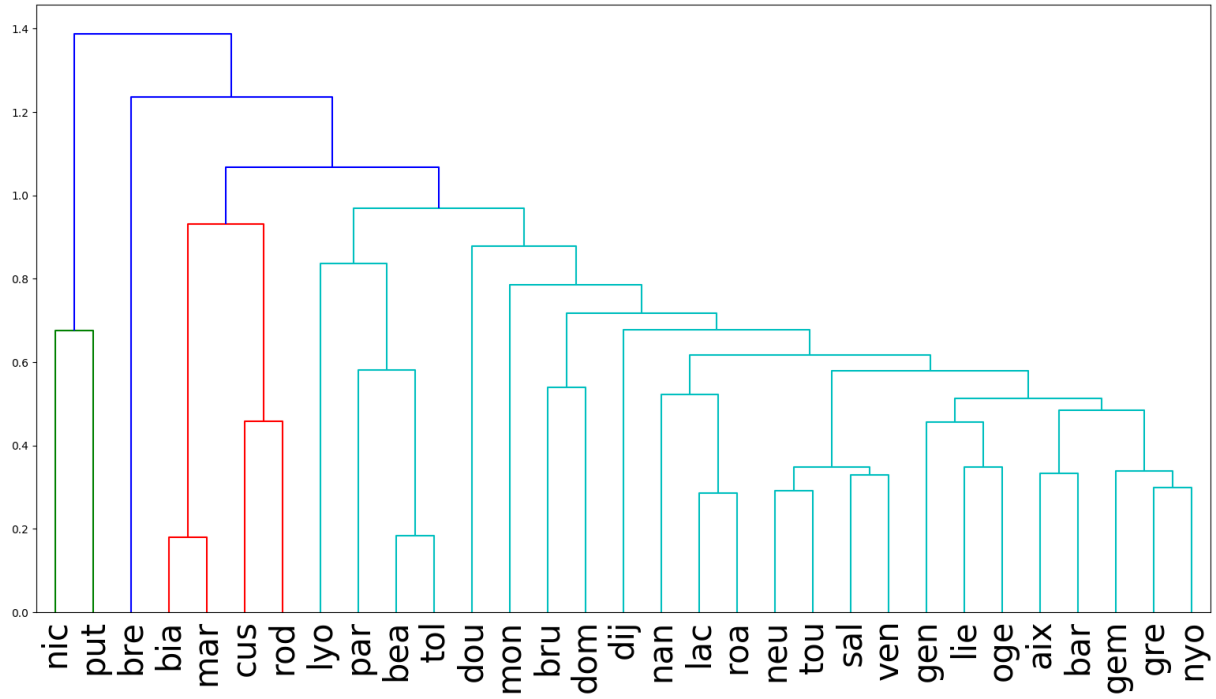
Figure 4: DET curve plot for prosodic features



3.4. Analysis

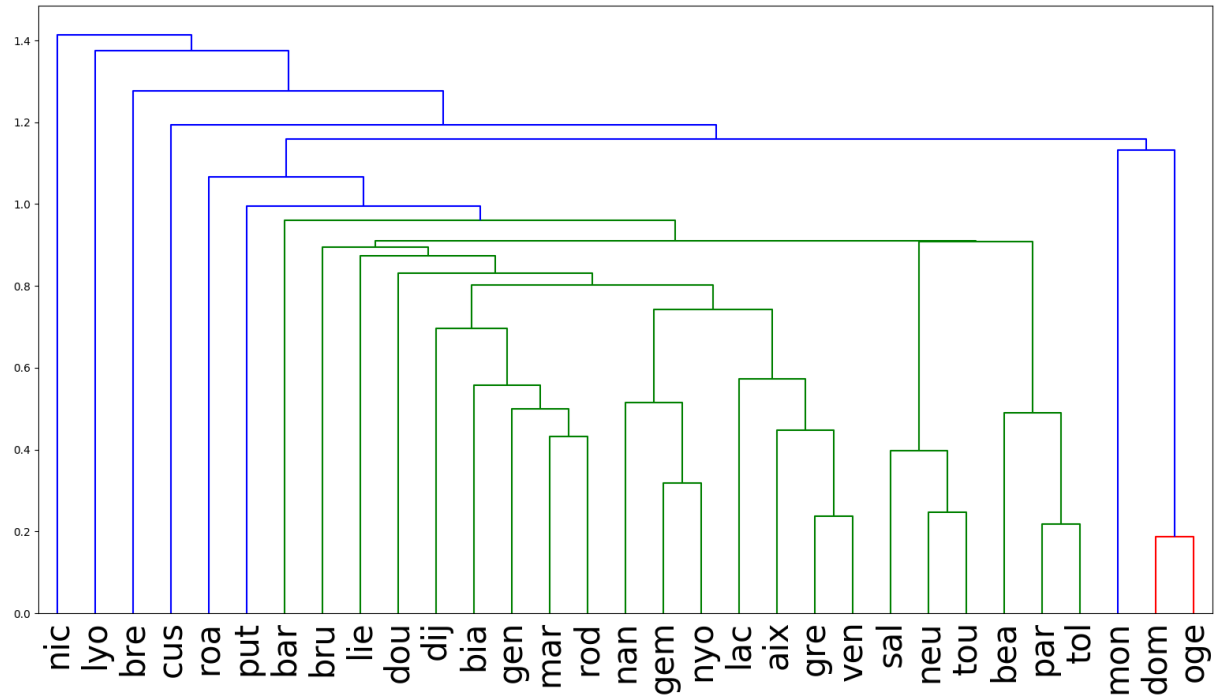
We have seen the accent detection performances for the different feature sets. Now we would like to see if the system errors are coherent with reality, i.e. whether it tends to confuse accents that are geographically or perceptively close to each other or not. Using the same method as Woerhling & Boula de Mareüil (2006), the confusion matrix for each experiment was used to do a hierarchical clustering of the regional accents. We used a complete link HAC algorithm and euclidian distance as distance function.

Figure 5: HAC for acoustic features



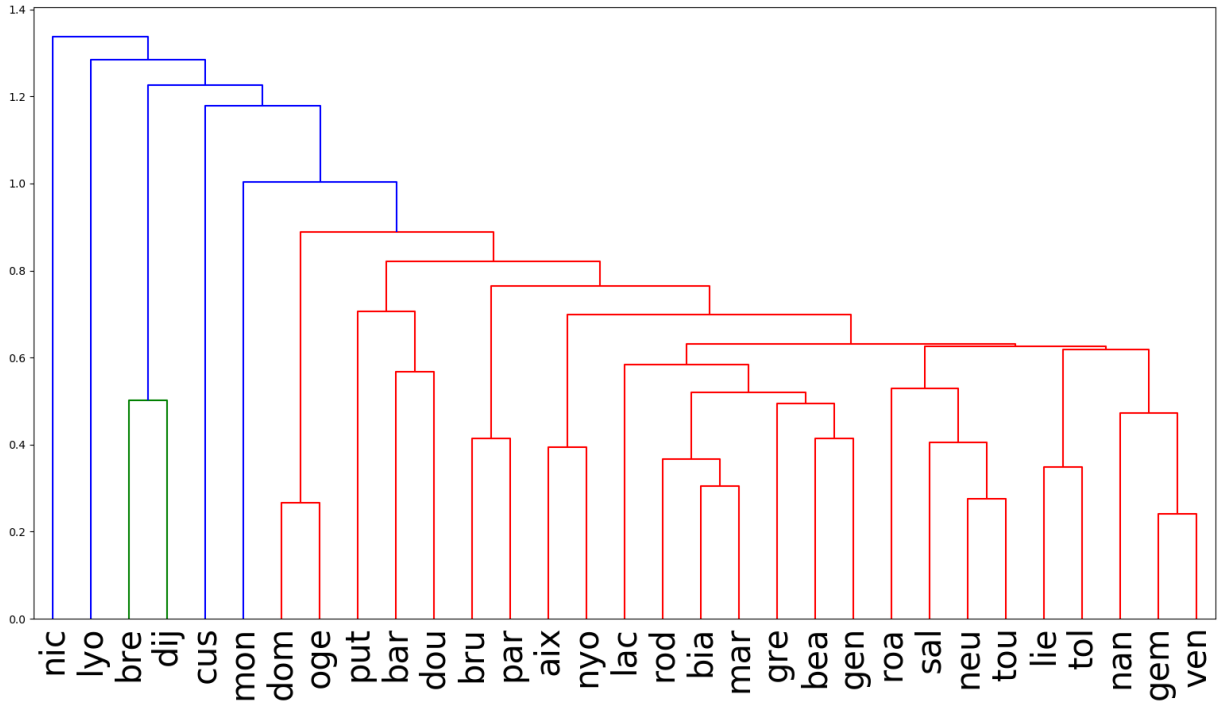
The red clustered comprised of Biarritz, Marseille, Cussac and Rodez is linguistically pertinent as all these investigation points are part of Southern French. Toulouse is also quite close to the Béarn. However the rest of the clustering does not seem to reflect a linguistic reality.

Figure 6: HAC for acoustic and pitch features



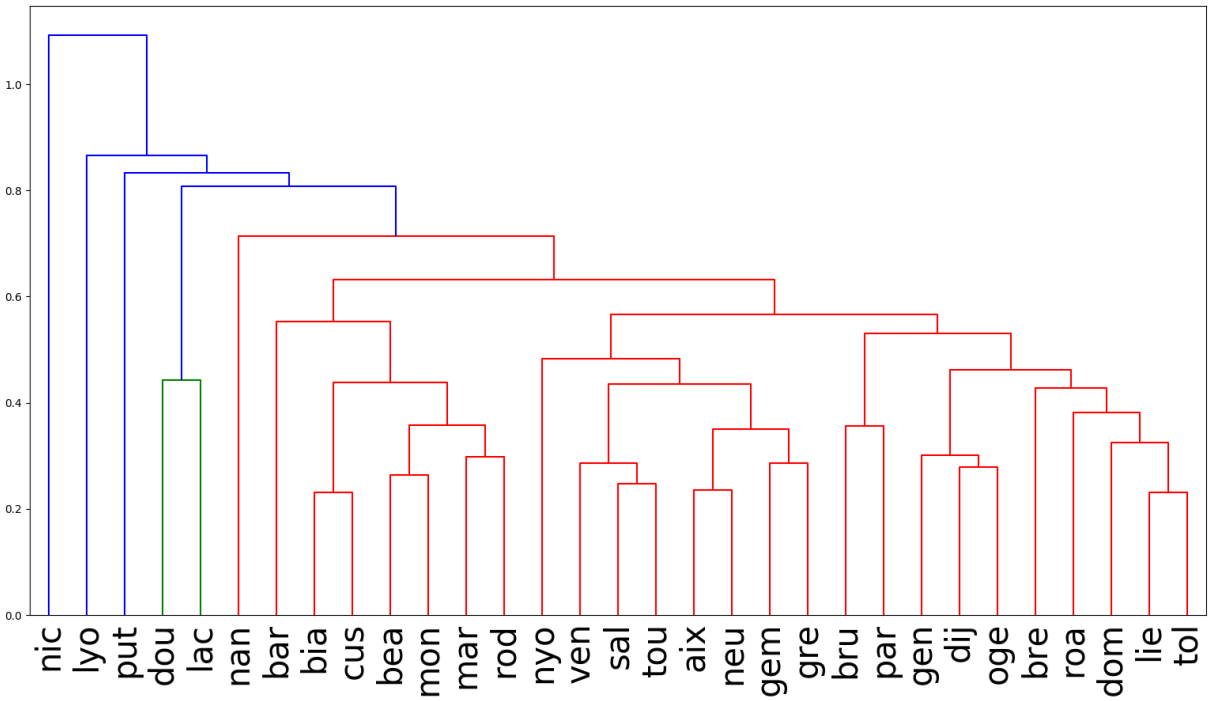
With the exception of the Marseille-Rodez cluster, the clustering obtained with the addition of pitch features does not seem better than the one with only acoustic features.

Figure 7: HAC for acoustic, pitch and RFC features



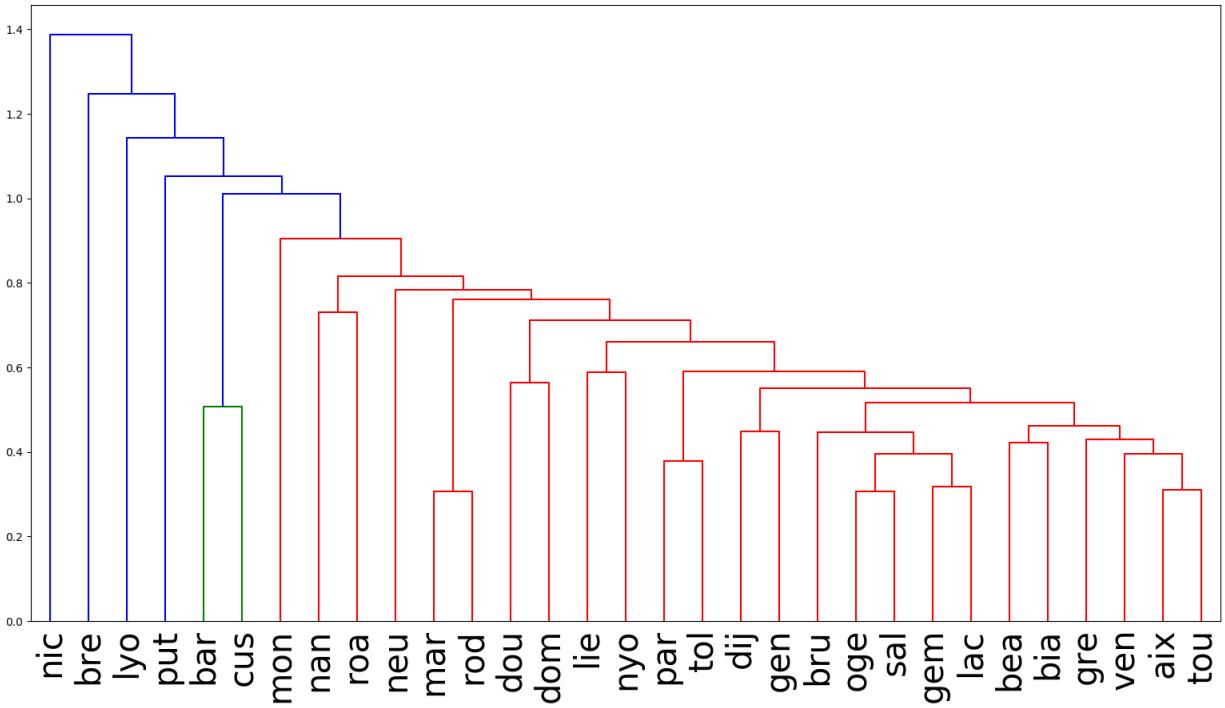
As previously, the Rodez-Marseille-Biarritz cluster is linguistically pertinent. The same goes for the Brunoy-Paris cluster.

Figure 8: HAC for acoustic, pitch and statistical intonation features



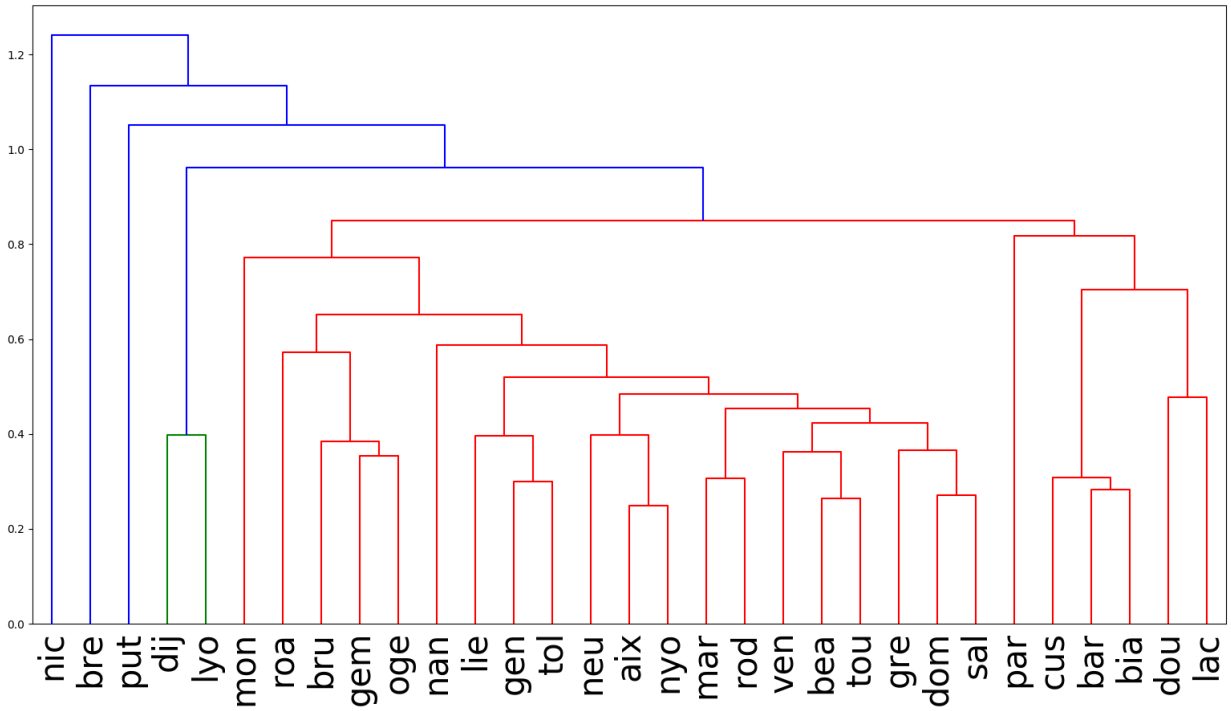
With the statistical intonation features, new pertinent clusters can be seen in addition to Brunoy-Paris, Marseille-Rodez and Biarritz-Cussac: Dijon-Ogéviller and Douzens-Lacaune.

Figure 9: HAC for acoustic, pitch and rhythm features



With the exception of the Marseille-Rodez and Nantes-Roanne clusters, none of the clusters seem to be linguistically pertinent.

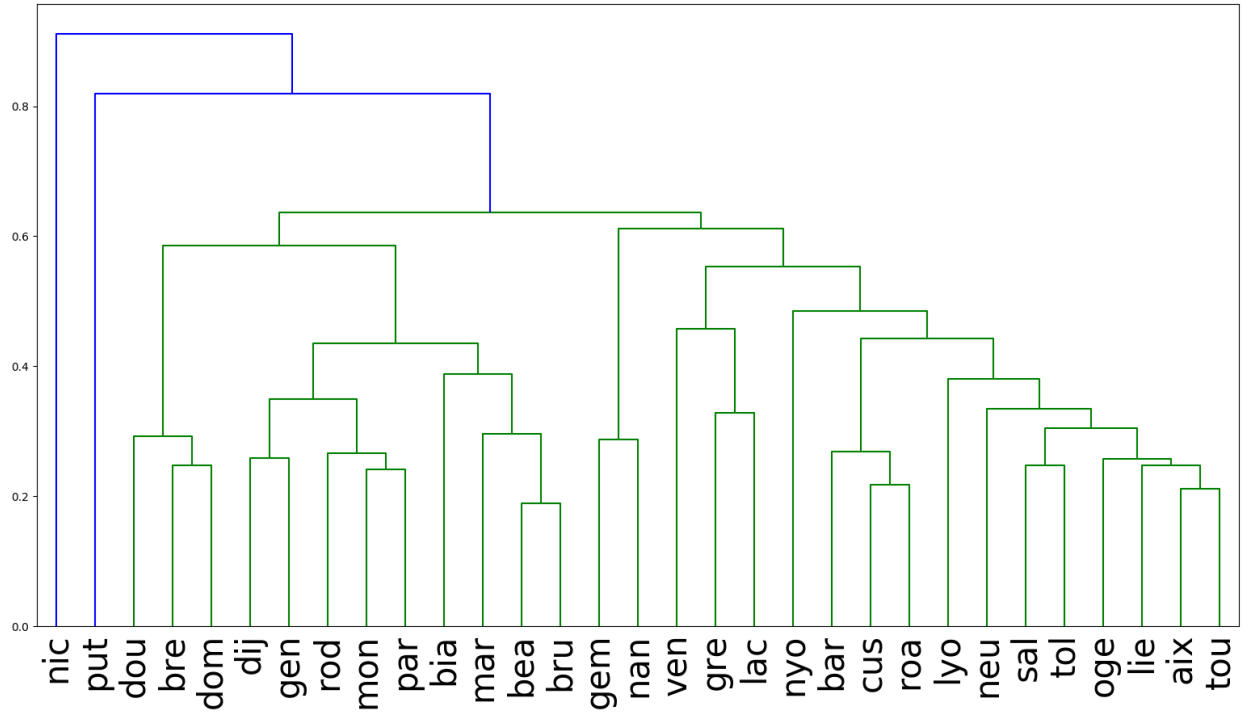
Figure 10: HAC for acoustic, pitch, RFC and rhythm features



When using RFC features and rhythm features at the same time, the clustering becomes more pertinent than with just one set of prosodic features. Interesting clusters can be seen such as Dijon-

Lyon, Marseille-Rodez, Neuchâtel-Aix-Nyon and Cussac-Bar-Biarritz-Douzens-Lacaune. For the last two, there is an investigation point that is an outsider (Aix and Bar-sur-Aube). There is a Mediterranean substrate in Swiss French and so having Aix in a Swiss French cluster is not completely aberrant. Concerning Bar-sur-Aube, although it is closer to Eastern French it shares common characteristics with Southern French such as fewer schwa elisions.

Figure 11: HAC for acoustic, pitch, statistical intonation and rhythm features



The addition of statistical intonation features and rhythm features gave the worst results however the clustering seems to be quite pertinent with clusters such as Brécey-Domfrontais, Montreuil-Paris and Salles-Toulouse.

In conclusion acoustic and pitch features gave the best results but the clustering was not satisfying. On the contrary intonation and rhythm features did not improve accent detection performances but gave some of the best clusterings.

4. Second experiment

We conducted a second experiment by dividing the 31 investigation points into 5 “global” accents: Northern French, Southern French, Eastern French, Belgian French and Swiss French.

Table 12: Global regional accents

Global regional accent	Investigation points
Belgian French (bel)	Gembloux, Liège, Tournai
Eastern French (est)	Bar-sur-Aube, Ogéviller
Northern French (nor)	Brécey, Brunoy, Dijon, Domfrontais, Grenoble, Lyon, Montreuil, Nantes, Paris, Puteaux-Courbevoie, Roanne, Vendée

Southern French (sud)	Aix-Marseille, Béarn, Biarritz, Cussac, Douzens, Lacaune, Marseille-Centre, Nice, Rodez, Salles-Curan, Toulouse
Swiss French (sui)	Genève, Neuchâtel, Nyon

The test and train datasets were the same as the first experiment.

4.1. Results

Here are the results obtained with the same sets of features as the first experiment:

Table 13: Results for acoustic features

	3 seconds	10 seconds	30 seconds	Total
Pmiss	0.6049	0.5615	0.5063	0.5870
Pfa	0.1512	0.1404	0.1266	0.1467
Cost	0.3781	0.3510	0.3164	0.3669

Table 14: Results for acoustic and pitch features

	3 seconds	10 seconds	30 seconds	Total
Pmiss	0.5520	0.4505	0.4767	0.5213
Pfa	0.1380	0.1126	0.1192	0.1303
Cost	0.3450	0.2816	0.2980	0.3258

Table 15: Results for acoustic, pitch, rhythm and statistical intonation features

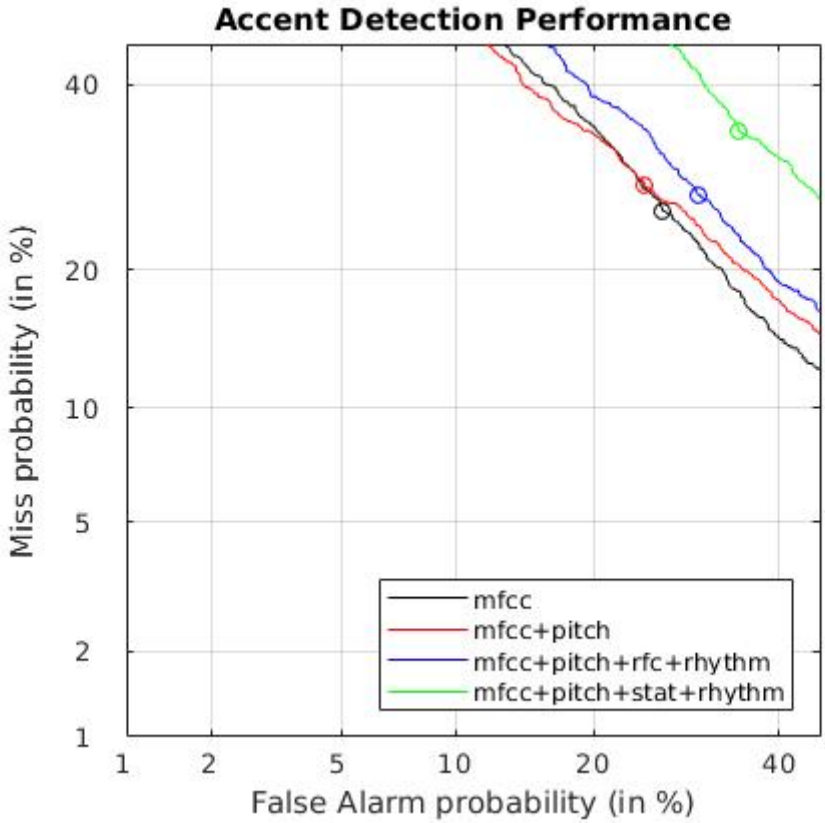
	3 seconds	10 seconds	30 seconds	Total
Pmiss	0.7277	0.7033	0.6476	0.7139
Pfa	0.1819	0.1758	0.1619	0.1785
Cost	0.4548	0.4396	0.4047	0.4462

Table 16: Results for acoustic, pitch, rhythm and RFC intonation features

	3 seconds	10 seconds	30 seconds	Total
Pmiss	0.6257	0.5576	0.5804	0.6057
Pfa	0.1564	0.1394	0.1451	0.1514
Cost	0.3911	0.3485	0.3628	0.3786

As for the first experiment, adding pitch features improves the accent detection performance compared to just acoustic features although prosodic features degrade performances.

Figure 12: DET curve plot for global accents

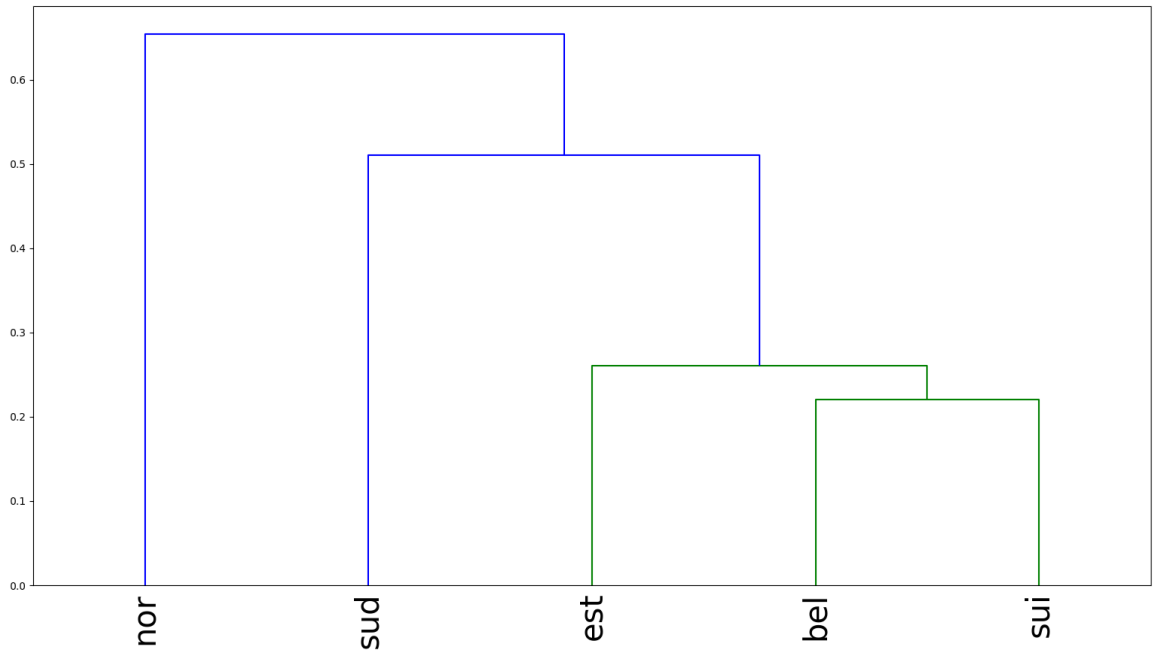


Although acoustic and pitch features combined gave better results than just acoustic features, the DET curves favors slightly acoustic features.

4.2. Analysis

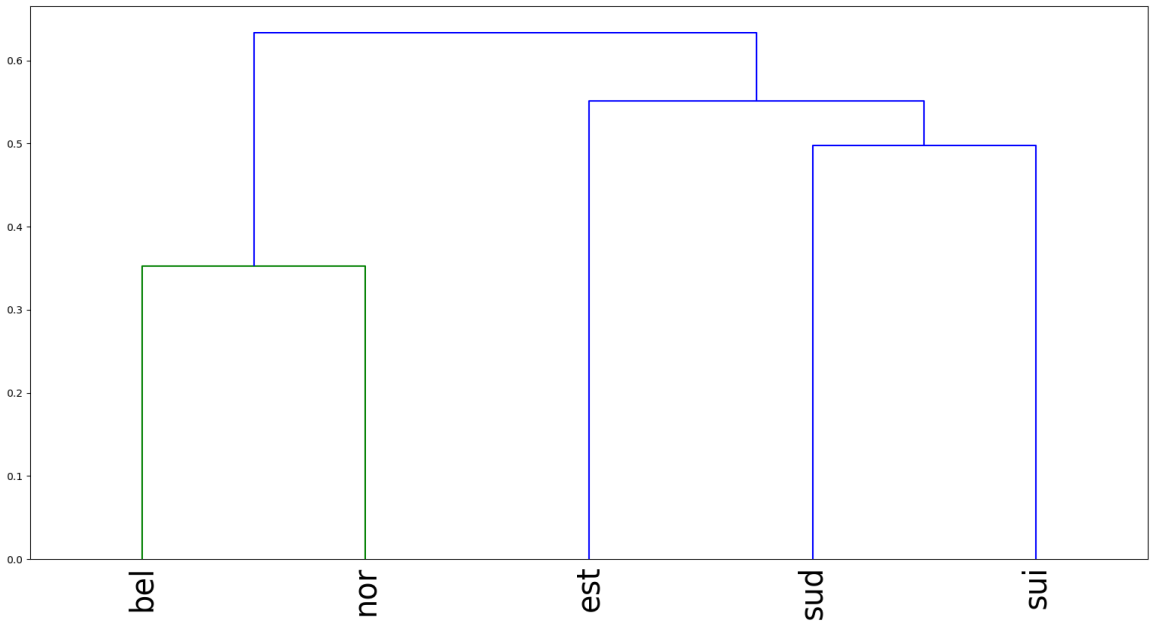
To see the significance of the prosodic features selected for the 31 investigations points in regards to the 5 global accents, a one-way ANOVA was conducted once again. All features were found to be significant except for pos_norm (p=0.806). Kurtosis, while significant, had a higher p-value compared to the 31 investigation points: p=0.0128. This may explain the important degradation in performances with the statistical intonation model.

Figure 13: HAC for acoustic features



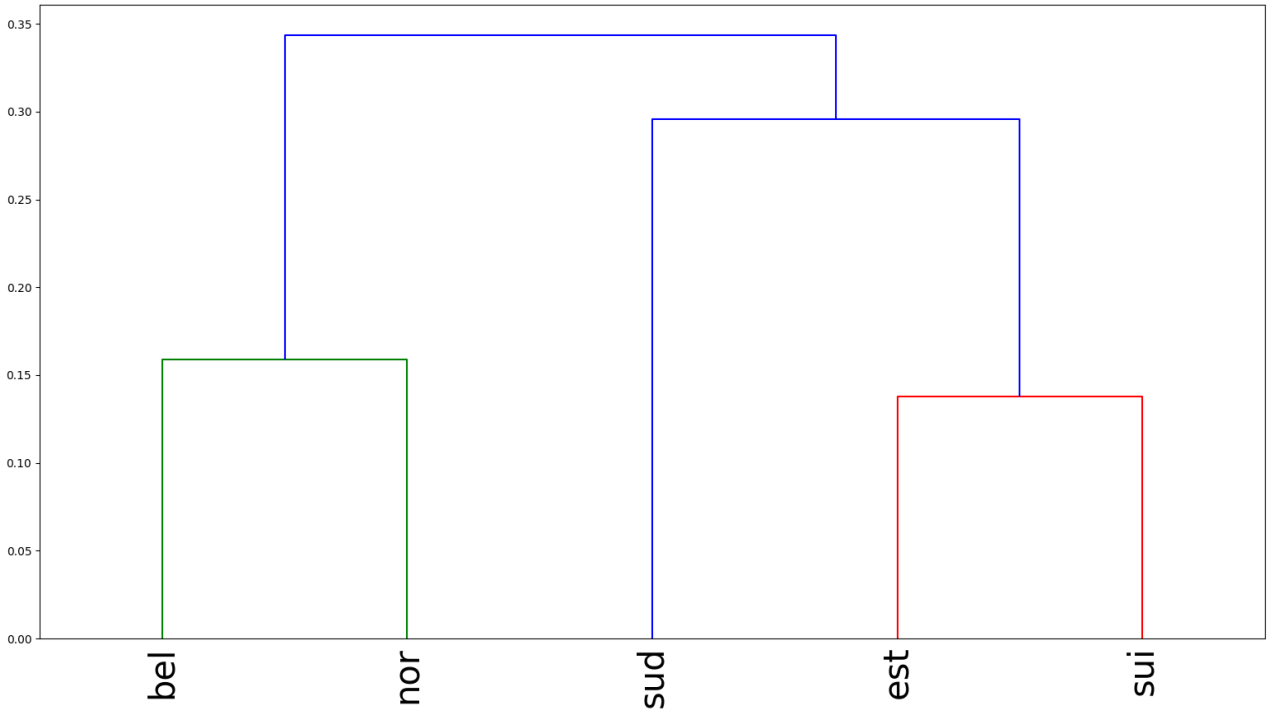
The hierarchical clustering for acoustic features gave results similar to Woerhling (2009) in that Eastern French, Belgian French and Swiss French are part of the same cluster. These three accents are then merged with Southern French. This may be due to the fact that all four accents tend to have long vowels: Southern French lengthens vowels before a nasal and there is still a length distinction in Eastern, Belgian and Swiss French (Woerhling 2009).

Figure 14: HAC for acoustic and pitch features



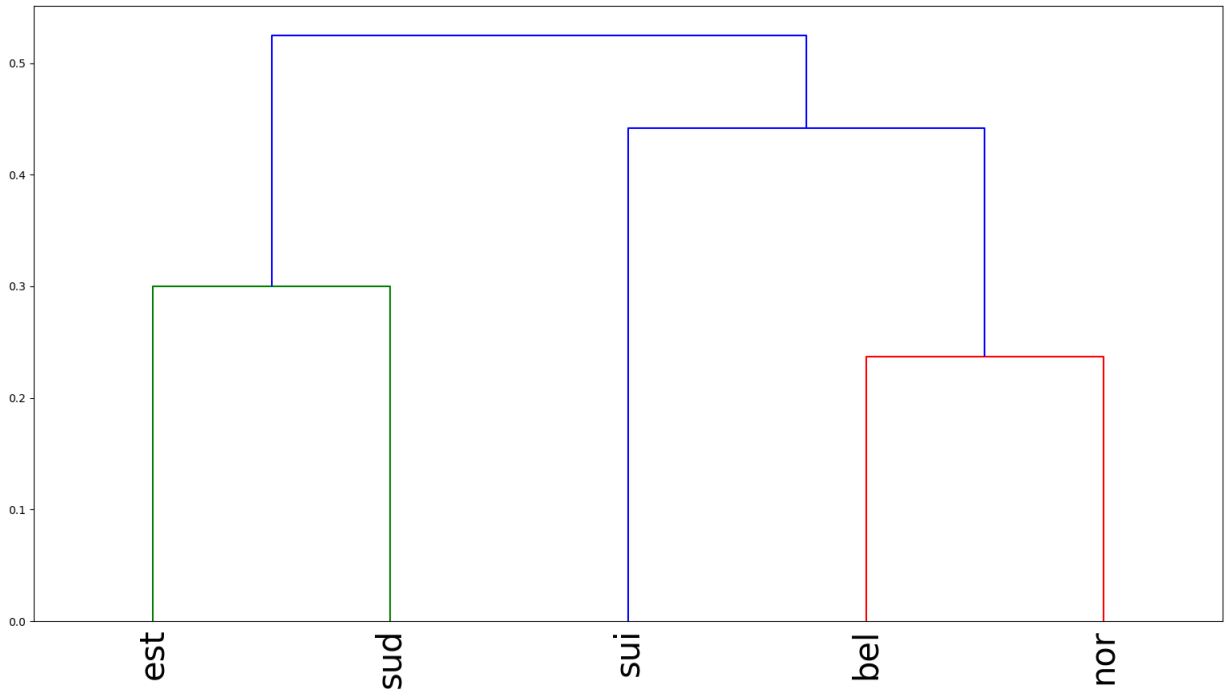
Contrary to the acoustic features, the clustering for acoustic and pitch features does not seem linguistically pertinent.

Figure 15: HAC for acoustic, pitch, rhythm and statistical intonation features



Adding rhythm and statistical intonation features gave a clustering quite similar to only acoustic and pitch features however Swiss French being merged directly with Eastern French is more pertinent comparing to the previous clustering.

Figure 16: HAC for acoustic, pitch, rhythm and RFC intonation features



As for statistical intonation features, the clustering with RFC intonation features is not more pertinent than with just acoustic features but is slightly better than with acoustic and pitch features.

Conclusion

We have achieved a 40% accuracy on regional accent identification with 31 investigation points in France, Belgium and Switzerland. Adding pitch features to the acoustic features helped improve performances up to 10 points however accent clustering was not linguistically pertinent. Although adding prosodic features did not improve performances, it improved accent clustering.

With 5 global regional accents we have achieved a 48% accuracy. The conclusions were the same as with the 31 investigation points: pitch improves accent detection but does not improve clustering and inversely prosodic features do not improve performances but may improve clustering.

Because of time limitations we ran the experiment using only one speaker for each investigation point as test data however to confirm the results of this study, the results of each set of features should be evaluated using cross-validation. As for the global regional accents, investigation points were divided intuitively and so a perceptive test should be carried out with native French speakers to determine an objective clustering.

The prosodic features used were extracted at the syllable level. It may interesting to extract features at the accentual phrase and utterance levels such as the number of syllables per accentual phrase, the mean syllable duration or a global pitch contour.

Bibliography

Régine André-Obrecht. 1988. A new statistical approach for automatic speech segmentation. *Transactions on Audio, Speech, and Signal Processing*, IEEE(36)1, 29-40.

Mohamad Hasan Bahari, Najim Dehak, Hugo van Hamme, Lukas Burget, Ahmed M. Ali & Jim Glass. 2014. Non-negative Factor Analysis of Gaussian Mixture Model Weight Adaptation for Language and Dialect Recognition. *IEEE/ACM transactions on audio, speech, and language processing* (22)7, 1117–1129.

W.J. Barry, .C.E. Hoequist & F.J. Nolan. 1989. An approach to the problem of regional accent in automatic speech recognition. *Computer Speech and Language*(1989) 3, 355-366.

Philippe Boula de Mareüil, Bianca Vieru-Dimulescu, Cécile Woehrling & Martine Adda-Decker. 2008. Accents étrangers et régionaux en français. Caractérisation et identification. *Traitement Automatique des Langues* 49(3), 135–162.

Jacques Durand, Bernard Laks & Chantal Lyche. 2002. La phonologie du français contemporain: usages, variétés et structure. In C. Pusch & W. Raible (eds.). *Romanistische Korpuslinguistik-Korpora und gesprochene Sprache/Romance Corpus Linguistics - Corpora and Spoken Language*. Tübingen: Gunter Narr Verlag, pp. 93-106.

Jacques Durand, Bernard Laks & Chantal Lyche. 2009. Le projet PFC: une source de données primaires structurées. In J. Durand, B. Laks & C. Lyche (eds). *Phonologie, variation et accents du français*. Paris: Hermès. pp. 19-61.

Jérôme Farinas. 2002. *Une modélisation automatique du rythme pour l'identification des langues*. Université Paul Sabatier.

Emmanuel Ferragne. 2008. *Étude phonétique des dialectes modernes de l'anglais des Îles Britanniques : vers l'identification automatique du dialecte*. Université Lumière Lyon II.

Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal & Sanjeev Khudanpur. 2014. A Pitch Extraction Algorithm Tuned for Automatic Speech Recognition. *ICASSP 2014*.

Alexandros Lazaridis, Elie Khoury, Jean-Philippe Goldman, Mathieu Avanzi, Sébastien Marcel & Philip Garner. 2014. Swiss French regional accent identification. *Odyssey*.

Piet Mertens. 2004. Un outil pour la transcription de la prosodie dans les corpus oraux. *Traitement Automatique des Langues* 45 (2), 109-130.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer & Karel Vesely. 2011. The Kaldi Speech Recognition Toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.*

Jean-Luc Rouas. 2005. *Caractérisation et identification automatique des langues*. Université Paul Sabatier.

Paul A. Taylor. 1992. *A Phonetic Model of English Intonation*. University of Edinburgh.

Cécile Woehrling. 2009. *Accents régionaux en français : perception, analyse et modélisation à partir de grands corpus*. Université Paris Sud - Paris XI.

Cécile Woehrling & Philippe Boula de Mareüil. 2006. Identification d'accents régionaux en français : perception et analyse. *Revue PArole* (37), 25–65.

Marc A. Zissman, Terry P. Gleason, Deborah M. Rekart & Beth L. Losiewicz. 1996. Automatic dialect identification of extemporaneous, conversational, Latin American Spanish speech. *Proceedings of ICASSP'96* (2). 777-780.