Towards Interactive Annotation for Hesitation in Conversational Speech

Jane Wottawa, Marie Tahon, Apolline Marin, Nicolas Audibert

Towards Interactive Annotation for Hesitation in Conversational Speech

Jane Wottawa[1,2], Marie Tahon[1], Apolline Marin[2], Nicolas Audibert[2]
[1]LIUM / Le Mans Université
[2]LPP, UMR 7018 CNRS - U. Paris 3 / Sorbonne Nouvelle
{jane.wottawa, marie.tahon}@univ-lemans.fr
apo.marin@bbox.fr, nicolas.audibert@sorbonne-nouvelle.fr

Abstract

Manual annotation of speech corpora is costly in both human resources and time. Furthermore, recognizing affects in spontaneous, non acted speech presents a challenge for humans and machines. The aim of the present study is to automatize the labeling of hesitant speech as a marker of expressed uncertainty. That is why, the NCCFr-corpus was manually annotated for DEGREE OF HESITATION on a continuous scale between -3 and 3 and the affective dimensions ACTIVATION, VALENCE AND CONTROL. In total, 5834 chunks of the NCCFr-corpus were manually annotated. Acoustic analyses were carried out based on these annotations. Furthermore, regression models were trained in order to allow automatic prediction of hesitation for speech chunks that do not have a manual annotation. Preliminary results show that the number of filled pauses as well as vowel duration increase with the degree of hesitation, and that automatic prediction of the hesitation degree reaches encouraging RMSE results of 1.6.

Keywords: Conversational speech, acoustic analyses, affective dimensions, hesitation, regression models.

1. Introduction

To better study hesitation in spontaneous conversational speech, large amounts of annotated data is required. One possibility to speed up the annotation process is to help the annotator by giving some label or value suggestions retrieved automatically (Marinelli et al., 2019). The second way is to reduce the number of segments to annotate by selecting only relevant data (Fallgren et al., 2019). Interactive annotation aims at combining both aspects to reach a good quality of annotation. An active learning strategy is usually included in the process in order to quickly reach a good quality of annotation suggestions by soliciting feedback from users if necessary (Klie et al., 2018). The quality of annotations can be validated using objective and subjective metrics obtained on a test data set, which represents a compromise between human perception and machine consistency.

For several years, affect in speech has been encoded using discrete categories such as *anger*, *sadness* or *neutral speech*. However, in many recent papers, researchers preferred using affective dimensions. Firstly theorized by Russel for emotional faces (Russel, 1997), the communication of affect can be seen as having three major dimensions of connotative meaning: *arousal* (activation), *pleasure* (valence) and *power* (dominance or control). In the field of affective computing, mainly activation and valence dimensions are used, and their predictions from speech is considered as a regression problem (Wöllmer et al., 2008). In more recent years, most of the neural prediction systems developed for affective computing, used convolutional or recurrent networks to predict activation and valence (Schmitt et al., 2019). The *control* (or dominance) dimension is also very important, especially in the context of conversational speech between humans or in human-robot interactions. Unfortunately only few spontaneous speech databases were annotated with this dimension (Tahon et al., 2010).

Most recent studies have developed machine learning systems able to predict emotional dimensions jointly (Parthasarathy and Busso, 2017), but very few of them went deeper in the acoustic analysis of these dimension.

In this paper, we focus on hesitation prediction in conversational speech. More precisely, we propose to investigate a continuum between hesitation and self-confidence, which is very close to the control dimension defined by (Scherer, 2005). As far as we know, the present paper is the first to tackle the issue of automatic continuous hesitation prediction.

Numerous studies on hesitation in spontaneous speech have been focusing on the distribution and duration of silent and filled pauses, such as those listed in the classical article by (Maclay and Osgood, 1959). To a lesser extent, syllabic lengthening has also been outlined as a correlate of hesitation, for instance in studies on French spontaneous speech by (Duez, 2001) or (Campione and Véronis, 2005).

Regarding the effect of hesitation on fundamental frequency (f_0), a study on German spontaneous speech (Mixdorff and Pfitzinger, 2005) found no impact of hesitations marked by fillers on the overall f_0 pattern at the utterance level. However, a study relying on synthesized speech (Carlson et al., 2006) in Swedish showed a moderate effect of the f_0 slope on perceived hesitation, as well as a moderate effect of the insertion of creaky voice.

Other studies focusing on expression of hesitation in speech have found similar acoustic correlates, with an important weight on silent pauses and fillers, and slighter differences on f_0 and voice quality-related parameters. For instance in English elicited speech, (Pon-Barry and Shieber, 2011) found significant correlations of hesitation ratings with temporal parameters (both silent pauses and speech rate), and to a lesser extent with f_0 slope, range and minimum value.

To conclude, hesitant speech is acoustically characterized by silent pauses, filled pauses as well as hesitation lengthening. Linguistically, hesitant speech includes repetitions reaching from syllables to word groups and finally more or less complex auto-corrections. As a rule, these phenomenon appear frequently in spontaneous and conversational speech

whereas they are quite rare in prepared and read speech. In French, native speakers produce filled pauses usually as *euh* ([œ] or [ø]) (Campione and Véronis, 2005; Vasilescu et al., 2004). Hesitation lengthening can either occur on filled pauses or function words. French function words counting at least two syllables, show hesitant lengthening usually on the last syllable. In a word chunk, more than one word can be affected by hesitant lengthening (Candea, 2000).
In the following, the used speech resource is first described, followed by the applied annotation frame. Furthermore, acoustic analyses carried out on the annotated speech chunks with respect to hesitation are presented followed by the section about automatic hesitation prediction. The paper ends with some concluding remarks about the annotation frame and analyses.

2. Speech resource

The corpus used in our study is the Nijmegen corpus of casual French (NCCFr) (Torreira et al., 2010). The corpus comprises recordings of 46 French speakers living in the same geographic area and with similar educational backgrounds. Each recording contains conversations from two speakers. Conversation partners knew each other well and were recorded sitting at a table.
Every participant was equipped with a microphone and recorded separately from the other. Every mono-channel recording is about 90 min long. However, as both speakers sat at the same table, their respective microphones did not only capture their speech but also, at a lower intensity, the speech of their partners.
The recordings were manually transcribed according to the French orthography. The orthographic transcriptions have then been automatically aligned to the signal using a Kaldi-based model developed at LIUM. For technical reasons, the recordings of 14 speakers were discarded. The following annotation frame was applied to the remaining 32 speakers.

3. Segmentation and annotation frame

This section presents the annotation frame which was applied to the speech productions of 32 speakers. All manual annotations were performed by the third author.

3.1. Segmentation protocol

The segmentation for the audio signal has been performed in two different steps. First, a manual segmentation was carried out on two minutes of each recording. Silent and filled pauses, false starts, word repetitions but also articulatory noises such as tongue clicks, lip licking or sharp inspirations were annotated.
In order to speed up the process, we decided to perform an automatic segmentation on the recordings. Silent pauses with a length superior to 200 ms were used as boundaries between speech chunks, using the phonetic transcription retrieved automatically.

3.2. Annotation scheme

The annotation frame was elaborated over time and adapted to the recordings. The recordings were annotated for DEGREE OF HESITATION and AFFECT (valence, arousal, control). In order to address the degree of hesitation, a Likert scale was used reaching from -3 (self-confident) to +3 (very hesitant). This system allowed us to attribute a score to each speech chunk situated between two pauses even if no hesitation was perceived.
The *arousal* dimension refers to the intensity of the affect expressed by speech, it also refers to activation, or intensity dimensions. *Valence* refers to how positive or negative the subject feels (Pereira, 2000). *Control* relates to the degree of dominance or sense of control over the expressed affect, and helps distinguish emotions initiated by the subject from those elicited by the environment. The last dimension is considered as a strong social cue during interactions between humans. In order to create a homogeneous annotation frame, the three features were also evaluated using a five level Likert scales reaching from -2 (extremely negative/passive/uncontrolled) to 2 (extremely positive/active/controlled). In this system, zero is supposed to represent neutral speech segments.
The perceptive and objective evaluation of affective dimensions is currently work in progress.

3.3. Corpus summary based on the annotation

The distribution of chunk duration along the degree of hesitation is summarized in Table 1. In conversational speech, the neutral state is usually over represented. Surprisingly, we can observe that the degree of -1 reaches the highest number of chunks, even higher than the degree 0 which was supposed to be neutral. Based on this observation, we decided to create meta categories comprising *sure* (degree of hesitation: -3, -2), *neutral* (-1, 0) and *hesitant* (1, 2, 3). Moreover, we can notice that the mean chunk duration is the smallest for the -1 degree, and we will see in the next section (sec. 4.), that this degree often behaves differently from the others.

Degree of Hesitation	Duration (s) mean	Duration (s) STD	#chunks (%)
-3	1.53	0.85	232
-2	1.49	0.97	333
-1	0.74	0.73	2 474
0	1.42	1.22	1 354
1	1.70	1.24	700
2	1.62	1.16	468
3	1.57	0.88	273
Total	1.20	1.07	5 834

Table 1: Distribution of chunks for each degree of hesitation: mean and standard deviation (STD) of chunk duration and number of chunks. Only annotated chunks are summarized.

4. Acoustic analyses

This section presents the acoustic analyses carried out on the manually annotated NCCFr-corpus. The analyses are specifically based on the annotation frame presented in Section 3.

A total of 232 acoustic features were derived from both symbolic information encoded in the forced alignment

output and acoustic analyzes obtained using custom Praat (Boersma and Weenink, 2019) scripts. Measures of duration, fundamental frequency (hereafter f_0), intensity relative to the mean level within each sound file, harmonics-to-noise ratio and zero-crossing rate were extracted, as well as frequencies of the first three formants for oral vowels. To enable their comparison between chunks with various segmental contents, formant values in Bark scale were converted to distances to centroids. Distances to centroids were computed both on the whole vocalic space of each speaker as a global measure of dispersion/centralization and relatively to the centroid of each vowel category as a measure of within-vowel variability, as in (Audibert et al., 2015). To get features relative to each chunk, analyzes were first carried out at the phone level before computing descriptive statistics for all phones included in a chunk. Raw and relative counts were computed for symbolic information (phones identity and phonological features). For numeric values, the mean, standard deviation, median, and a set of percentiles (5%, 25%, 75% and 95%) as well as the minimum and maximum values were computed on the distribution of values over each chunk, for all phones and separately for vowels and consonants. In addition to those descriptive statistics, measures taking the time dimension into account were computed: slope of the regression over time on the whole chunk, relative value of the first and last phone, and relative time of the minimum and maximum value. For duration values, nPVI values (Grabe and Low, 2002) were also computed as a measure of local rhythmic variation.

In the following, the number of filled pauses according to degree of hesitation, mean vowel duration per chunk according to degree of hesitation, f_0 of the last vowel according to degree of hesitation, and the length of silent pauses according to the degree of hesitation of the following speech chunk are presented.

4.1. Occurrences of filled pauses

The presence of filled pauses is known to be linked to hesitation in speech. In French, filled pauses are usually uttered as *euh* [œ] or [ø]. For each degree of hesitation, all *euh* were extracted. A ratio of the number of *euh* occurrences by chunk was calculated for each degree of hesitation. Figure 1 summarizes the results. The number of *euh* per chunk increases the more hesitant the speech becomes, reaching from 0.07 to 5.6 on the extreme ends of the Likert scale.

Table 2 summarizes filled pauses duration according to degree of hesitation. For the -3 degree, only one occurrence of *euh* was present in our data that is why no standard deviation was calculated. Overall, the duration of filled pauses increases once speech was annotated as hesitant *i.e.*, the degrees of 1 up to 3. The duration seems similar for the degrees of -3 up to 0.

Both the number of occurrences as well as the duration of filled pauses seem to qualify hesitant speech. In our data, hesitant speech chunks present more and longer filled pauses than do the speech chunks rated as sure or neutral.

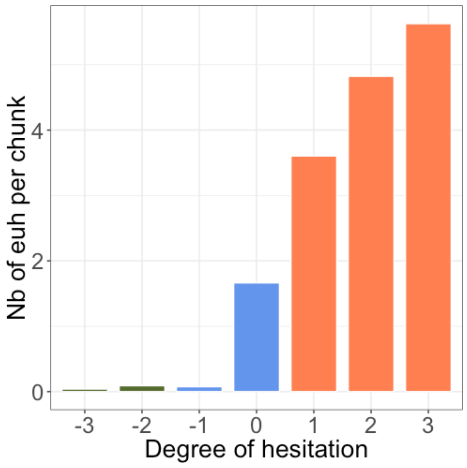

Figure 1: Normalized occurrences of *euh* according to hesitation degree. (green: sure, blue: neutral, orange: hesitant).

Degree of Hesitation	Duration of *euh* (s)	
	mean	STD
-3	0.18	not defined
-2	0.14	0.06
-1	0.16	0.10
0	0.14	0.08
1	0.20	0.11
2	0.27	0.13
3	0.47	0.26
Total	0.27	0.19

Table 2: Mean duration of filled pauses according to degree of hesitation. (-3/-2: sure, -1/0: neutral, 1/2/3: hesitant).

4.2. Articulation rate

Articulation rate is linked to the degree of hesitation. Hesitation in speech is known to be linked to verbal planning that can influence both lexical choices and syntactic planning. Other than the anticipation of linguistic elements, hesitation might be provoked by uncertainty of the communicated content. In both cases, linguistic planning and uncertainty, articulation rate decreases.

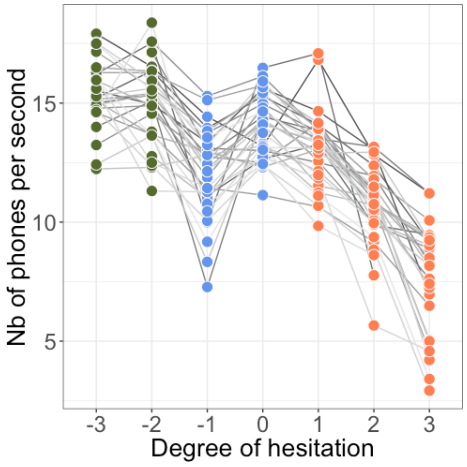

Figure 2: Articulation rate according to hesitation degree. Each set of points connected by lines represent a participant. (green: sure, blue: neutral, orange: hesitant)

An ANOVA with the within-subjects factor DEGREE OF HESITATION indicated that participants produce statistically different articulation rates for the seven degrees of hesitation ($F_{(6,\,176)} = 93.3, p < .001$). Figure 2 represents this result. Globally, articulation rate decreases the more the degree of hesitation increases. However, the -1 degree stands out in comparison with its environment. Articulation rate is globally lower than for the degrees -2 and 0. Global analyses (*i.e.*, Table 1) showed, that, compared to the other degrees, the -1 degree contains very short chunks which could explain the lower articulation rate. Shorter chunks contain less phonemes and are more likely to be interrupted by longer silent pauses. Both factors lead to a decreased articulation rate.

4.3. Vowel duration

Syllable lengthening is a strategy which allow to avoid filled or silent pauses. In French, syllable lengthening is achieved by producing longer vowels, especially in the final syllables of polysyllabic words. The mean vowel duration per chunk according to hesitation degree was analyzed. The results of an ANOVA with the within-subjects factor DEGREE OF HESITATION was run and indicated that the mean vowel duration per chunk varies for all speakers across all seven degrees of hesitation ($F_{(6,\,176)} = 84.1, p < .001$). Figure 3 illustrates this result. The figure shows a general increase of mean vowel duration with increasing hesitation. This trend was globally observed for all speakers.
Furthermore, VOWEL DURATION and ARTICULATION RATE are negatively correlated ($r = -0.65$). The longer the vowels, the slower the articulation rate. This result was expected as an increased vowel duration slows down the number of phones uttered per second.

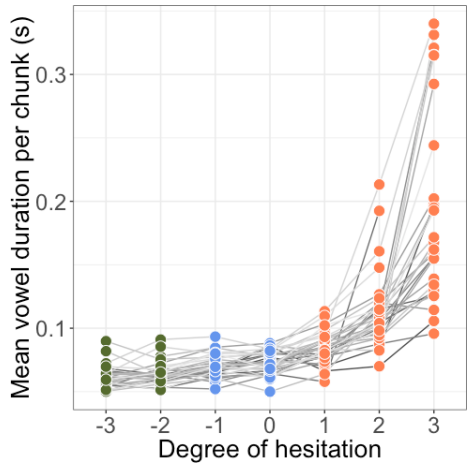

Figure 3: Mean vowel duration according to hesitation degree. Each set of points connected by lines represent a participant. (green: sure, blue: neutral, orange: hesitant)

4.4. f_0 of the last vowel in a chunk

The fundamental frequency can inform us about melody changes in speech. French is a syllabic language which marks focus rather by word order and lexical choices than melody. However, melody changes might be present at the end of utterances. That is why analyses of the f_0 on the last vowel of the chunks were carried out.

ANOVA with the within-subjects factor f_0 OF THE LAST VOWEL IN A CHUNK revealed that participants differ their f_0 according to the degree of hesitation individually ($F_{(6,\,172)} = 15.7, p < .001$). Figure 4 illustrates this result where two distinct ensembles can be observed. As the f_0 is strongly influenced by gender, the lower graphs, situated between 90 and 150 Hz, belong probably to the male speakers of the data set whereas the upper graphs, situated between 150 and 300 Hz, belong to the female speakers. Both gender groups show a similar f_0 pattern across the different degrees of hesitation. In French, hesitant speech seems to be linked to a lowering of the f_0, compared to neutral and certain speech.

Automatic (as well as manual) f_0 extraction can lead to erroneous values which could explain the outliers present in Figure 4 located at the 2 and 3 degrees.

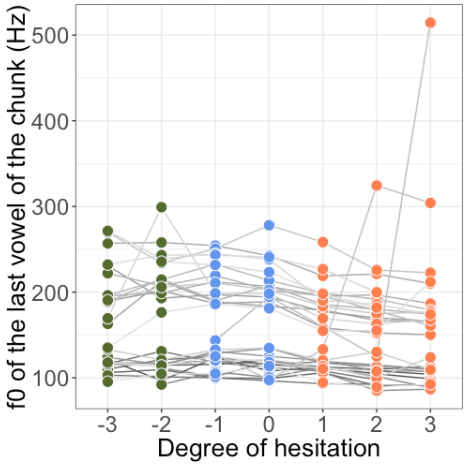

Figure 4: Mean f_0 of the last vowel of the chunk according to hesitation degree. Each set of points connected by lines represent a participant. (green: sure, blue: neutral, orange: hesitant)

4.5. Duration of silent pauses

Furthermore, the duration of the silent pauses preceding the chunks were analyzed in order to identify whether the length of silent pauses predicts the degree of hesitation of the following chunk. The analyses remained inconclusive. The duration of pauses does not seem to be linked to the degree of hesitation of the following chunk.

The acoustic analyses have shown that hesitant speech is linked to an increased number of filled pauses, a decrease in articulation rate which is correlated to an increasing vowel duration and a lowering of f_0 on the last vowel of speech chunks. However, analyses of all speakers pooled were not conclusive for the acoustic analyses of ARTICULATION RATE, VOWEL DURATION, and f_0 OF THE LAST VOWEL IN A CHUNK. These results indicate that speakers have individual strategies to express hesitation. Their strategies depend, among other things, on their individual articulation rate, vowel duration and f_0. That is why acoustic analyses for hesitant speech rarely show robust results that account for whole speaker groups.

5. Automatic prediction of hesitation

Our global objective is to develop an interactive annotation protocol using active learning (AL) for a better automatic annotation of the whole corpus. In order to achieve this aim, we propose preliminary experiments on automatic prediction of the degree of hesitation. This task can be seen as a regression problem, and in the following, different regressive models will be evaluated.

5.1. Protocol description

A selection function, which is included in the AL process, has to select the instances that are likely to improve the model's performances. Usually the selection function relies on probability or uncertainty of the prediction. Because not all regressive models do generate interpretable probability output functions, we decided to use the Query By Committee (QBC) protocol (Settles, 2009). In this protocol, different regressive models are considered: when results are consistent across models (*i.e.*, low standard deviation), we believe that the prediction is robust.
In this paper, we do not present the final AL results, but preliminary results on hesitation prediction. We have tested six different regressive models (section 5.2.), with two different feature sets (section 5.3.).
Our data is split in 3 subsets as summarized in Table 3: two training sets and a test set. A large and a small training set were defined in order to estimate the impact of data quantity. Each train set is split randomly into a sub-training set (80%) and a development set (20%). The development set is used for parameter optimization and feature selection.

- *testSet*: 10 first minutes of 2 speakers (1M, 1F),
- *SmallSet*: 2 first minutes of 30 speakers,
- *LargeSet*: 10 first minutes of 30 speakers.

This partition ensures us to be speaker-independent in the evaluation of our models.

Subset	Train	Dev	TotTrain	Test
SmallSet	892	224	1116	360
LargeSet	4379	1095	5474	360

Table 3: Number of chunks of each train and test sets.

5.2. Regression models

Six different regression models are explored: three Support Vector Machines for Regression (SVR) with different kernels (gaussian, polynomial and linear), a classical linear regression (LinReg), Lasso, and Ridge regression algorithms. SVM are known to be a robust approach for affective state modeling, providing that parameter optimization is systematically done. Linear regression is one of the most simple algorithms for regression tasks, it has the advantage of being fast and parameter free. The inconvenient is that the algorithm includes all features at the same level which leads to overfitting. Ridge and Lasso regression are powerful techniques generally used for creating parsimonious models in the presence of a large number of features, consequently they avoid overfitting. Both of them are also regularization techniques since they penalize features while minimizing the error between predicted and actual labels. Lasso uses a L1 regularization, while Ridge uses L2 regularization. Consequently Lasso and Ridge can also be used as feature ranking approaches.
We know that Neural Networks are state of the art models, however such architectures are not convenient for our task: they require a lot of data, take a long time to run efficient models, and model adaptations (using transfer learning approaches) also require a big amount of data.
Our regression problem aims at predicting isolated and independent values for each speech segment. Therefore, metrics including correlation are not appropriate (*e.g.* correlation coefficient, concordance correlation coefficient). The metric that best fits with our problem is the root mean square error defined as the average squared difference between target (y_k) and predicted ($\hat{y_k}$) values of segment k over all segments (see equation 1).

$$\texttt{RMSE} = \sum_{k=1}^{N} (\hat{y_k} - y_k)^2 \quad (1)$$

5.3. Acoustic feature and feature selection

In this first study on hesitation, we decided to train our models on acoustic features only. Two sets are investigated:

- *acousFeat*: a 232 acoustic feature set described in 4.
- *melFeat*: 4×20 cepstral values per segment: 20 cepstral coefficients (MFCCs) are extracted at the frame level (nfft=512 and step=128). Mean and relative standard deviations (std/mean) are computed at the segment level on MFCCs and their first derivative ΔMFCCs.

Articulation rate has not been included in *acousFeat* because of its high correlation with vowel duration.
All features were normalized on the subset (train, development or test) so that the mean is 0 and the standard deviation is 1. We first trained regressive models separately with the whole set. This process allowed us to a) optimize model parameters for the given task and b) select acoustic features using a forward feature selection approach (Tahon et al., 2018). Then parameters were optimized again using the selected feature set.
The automatic feature selection processed on *acousFeat* allowed to retrieve the prosodic features analyzed in section 4..
At the chunk level, among the mostly selected features (6 models $\times$ 2 data sets Small and Large), we found the following features (with their selection frequency):

- maximum vowel duration (10), mean duration of the first phone (7),
- amount of phonemes: [ø] and [œ] (10)
- f_0 of the last vowel (3)
- harmonic-to-noise maximum ratio on vowels (4)
- fusion (10)

The last fusion parameter is very interesting. The feature is directly linked to the segmentation process. Indeed, we created this parameter to identify chunks for which manual and automatic segmentation differed. fusion=1 means that there were two manual chunks for only one automatic chunk. The authors are aware that this feature can not be extracted directly from the signal. But its high rank among the selected features shows that human perception has a great impact on the automatic segmentation of hesitation chunks.

5.4. Results

Model	SmallSet		LargeSet	
	Dev	Test	Dev	Test
svr-rbf	0.87	1.55	0.87	1.58
svr-poly	1.49	2.41	1.85	2.37
svr-lin	0.96	1.64	1.00	1.69
lin-reg	0.92	1.68	1.02	1.77
lasso	1.17	1.64	1.17	1.67
ridge	0.94	1.62	1.03	1.75
uQBC	-	1.64	-	1.68
wQBC	-	1.62	-	1.67

Table 4: Automatic prediction of hesitation with optimized parameters and *acousFeat* selected features. RMSE obtained for each regressive model on both development and test sets. RMSE obtained using the Query By Committee unweighted (uQBC) and weighted (wQBC) methods.

Table 4 summarizes the RMSE scores obtained on both sets development and test separating the results obtained on the *SmallSet* (2 min) and the *LargeSet* (10 min). The reader should be aware that the results obtained on the development set are speaker dependent, while the results on the test set are speaker independent. This assumption could explain the huge drop in performance between development and test sets.

The values obtained on *melFeat* are much lower – when the six models are trained on the *SmallSet*, we found an average RMSE of 1.37 over the development set – than the ones obtained using *acousFeat* (average RMSE is 1.06). That is why, only regression results obtained with *acousFeat* are reported in Table 4.

The Query By Committee value is also given as the mean value over the 6 predicted values per chunk. The mean is either unweighted (uQBC) or weighted (wQBC). Weights are given by the inverse performances reached by models on the development test: a good performance (a low RMSE value) leads to a high weight. As expected, wQBC seems to be a better way to merge model predictions.

The best model is svr-rbf, while the least efficient is svr-poly, reminding us that svr kernels must be carefully chosen according to the task and data. Lasso models are also interesting as the poor score obtained on dev (1.17/1.17) has not a huge impact on test (1.64/1.67). Interestingly, the common idea that more data leads to better performances is not validated in our study. Performances are lower on the *LargeSet* (wQBC = 1.67) than on the *SmallSet* (wQBC = 1.62). Which means that with additional data come not only potentially new examples that the models can learn but also more noise. We hypothesize that this noise might be linked the heterogeneity of the annotations across speakers and time.

These results underline that selecting the chunks that should be annotated could help to improve modeling the degree of hesitation.

6. Conclusions and discussion

The present paper presented preliminary results on the segmentation and annotation performed on the NCCFr corpus in order to develop an interactive annotation protocol.

Segmentation was achieved through manual annotation of silent and filled pauses and articulatory noise of the first two minutes of 32 speakers' recordings. Based on this first segmentation, the recordings were then automatically segmented and aligned on a phoneme level.

The annotations of the corpus consisted in the manual attribution of a degree of hesitation to each speech chunk situated between silent pauses as well as the annotation of affective dimensions (activation, valence, control). A total of 5834 chunks were manually annotated.

Acoustic analyses indicated that speech with a high hesitation score contains a larger number of filled pauses, has a lower speech rate, longer vowel duration and a relatively low f_0 on the last vowel of a chunk. However, speakers have individual strategies to achieve these tendencies. Values obtained for the whole speaker group are not conclusive. Every speaker has an individual speaking style, hesitation can be expressed only in function of this individual speaking style. Thus all adaptations of the speaking styles to different degrees of hesitation are individual as well and cannot be summed up as a group mean. However, the tendencies of the individual changes remain similar across the group.

The data suggests that the annotator based her attribution of the degree of hesitation on the number of filled pauses present in the speakers' speech. Filled pauses are commonly associated with hesitant speech. This result is not surprising but could explain certain biases in the data set. That is why we are currently collecting further manual annotations from a pool of annotators in order to reduce individual annotation biases for chosen segments of the data.

With respect to the regression models, it is possible to predict a degree of hesitation. We have shown that prosodic features such as vowel duration, f_0, and articulation rate characterize the degree of hesitation.

This study validates the relevance of a new affective dimension which extremes are certain and extremely hesitant.

In our case, regression models for hesitation prediction do not benefit from the addition of more annotated data as noise and inconsistencies increase with a larger amount of data. Therefore there is room for improvement in the selection of relevant chunks that should be annotated, then included in the training process. In future work, we plan to study different strategies for active learning in the context of the development of interactive annotation tools.

Our current analyses of the corpus do not take into account the interactivity of our speakers. Throughout all analyses, speakers were considered independently from each other. In future work, we would like to analyze their interaction and the interplay of their interaction and the degree of hesitation attributed to their respective speech chunks.

7. Acknowledgements

This work was supported by the Labex EFL program (ANR-10-LABX-0083). We thank our colleague Antoine Laurent for the phonetic alignment tool he developed for our purpose.

8. Bibliographical References

Audibert, N., Fougeron, C., Gendrot, C., and Adda-Decker, M. (2015). Duration-vs. style-dependent vowel variation: A multiparametric investigation.

Boersma, P. and Weenink, D. (2019). Praat: Doing phonetics by computer [Computer program]. Version 6.0. 45.

Campione, E. and Véronis, J. (2005). Pauses and hesitations in French spontaneous speech. In Disfluency in Spontaneous Speech Workshop, pages 43–46.

Candea, M. (2000). Les euh et les allongements dits «d'hésitation»: deux phénomènes soumis à certaines contraintes en français oral non lu. *XXIIIèmes Journées d'étude sur la Parole*, pages 73–76.

Carlson, R., Gustafson, K., and Strangert, E. (2006). Cues for hesitation in speech synthesis. In Ninth International Conference on Spoken Language Processing.

Duez, D. (2001). Signification des hésitations dans la parole spontanée. *Revue parole*, pages 17–18.

Fallgren, P., Malisz, Z., and Edlund, J. (2019). How to Annotate 100 Hours in 45 Minutes. In Proc. Interspeech 2019, pages 341–345.

Grabe, E. and Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in laboratory phonology*, 7(515-546).

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In CoLing, pages 5–9, Santa Fe, New Mexico, USA, August.

Maclay, H. and Osgood, C. E. (1959). Hesitation phenomena in spontaneous english speech. *Word*, 15(1):19–44.

Marinelli, F., Cervone, A., Tortoreto, G., Stepanov, E. A., Fabbrizio, G. D., and Riccardi, G. (2019). Active Annotation: Bootstrapping Annotation Lexicon and Guidelines for Supervised NLU Learning. In Proc. Interspeech 2019, pages 574–578.

Mixdorff, H. and Pfitzinger, H. R. (2005). Analysing fundamental frequency contours and local speech rate in map task dialogs. *Speech Communication*, 46(3-4):310–325.

Parthasarathy, S. and Busso, C. (2017). Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning. In Interspeech 2017, pages 1103–1107. ISCA, August.

Pereira, C. (2000). Dimensions of Emotional Meaning in Speech. pages 25–28, Newcastle, Nothern Ireland, UK, September.

Pon-Barry, H. and Shieber, S. M. (2011). Recognizing uncertainty in speech. *EURASIP Journal on Advances in Signal Processing*, 2011(1):251753.

Russel, J. (1997). Reading emotions from and into faces: Resurrecting a dimensional-contextual perspective. In *The psychology of facial expression*. Cambridge University Press, U.K., pp. 295–360.

Scherer, K. R. (2005). What are emotions ? and how can they be measured ? SAGE Publications, chapter Social Science Informationpp. 695–729.

Schmitt, M., Cummins, N., and Schuller, B. W. (2019). Continuous Emotion Recognition in Speech — Do We Need Recurrence? In Interspeech 2019, pages 2808–2812, Graz, Austria, September. ISCA.

Settles, B. (2009). Active Learning Literature Survey. Technical report.

Tahon, M., Delaborde, A., Barras, C., and Devillers, L. (2010). A corpus for identification of speakers and their emotions. In Language Ressources and Evaluation Conference (LREC), Valleta, Malta.

Tahon, M., Lecorvé, G., and Lolive, D. (2018). Can we Generate Emotional Pronunciations for Expressive Speech Synthesis? *IEEE Transactions on Affective Computing*.

Torreira, F., Adda-Decker, M., and Ernestus, M. (2010). The Nijmegen corpus of casual French. *Speech Communication*, 52(3):201–212.

Vasilescu, I., Candea, M., and Adda-Decker, M. (2004). Hésitations autonomes dans 8 langues: une étude acoustique et perceptive. In Workshop MIDL04, Paris, France.

Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., and Cowie, R. (2008). Abandoning Emotion Classes – Towards Continuous Emotion Recognition with Modelling of Long-Range Dependencies. In Proc. Interspeech, pages 597–600.