Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 277-283 recibido 29-04-2005; aceptado 01-06-2005

Analysis of prosodic features: towards modelling of emotional and pragmatic attributes of speech *

Jordi Adell, Antonio Bonafonte
Universitat Politècnica de Catalunya
c/Jordi Girona 1-3 D5
08034-Barcelona
www.talp.upc.es

David Escudero
Universidad de Valladolid
Campus Miguel Delibes s/n
47011-Valladolid
www.infor.uva.es

Resumen: Aunque las tecnologías de voz mejoran de forma constante sus prestaciones, es necesario comprender los mecanismos utilizados en el habla para transmitir, además del léxico, otras informaciones como la emoción, actitud o estilos del hablante. En este trabajo nos hemos centrado en el estudio de la correlación de los parámetros básicos de la prosodia con características del tipo emocional y pragmático. Para ello, se han utilizado tres corpora: grabaciones de voz con emoción, lectura de una novela y sesiones del parlamento español. Basándonos en una taxonomía de emociones y modos del discurso, hemos planteado el estudio mediante una tarea de clasificación en base a las características prosódicas. Los resultados preliminares muestran que podemos identificar claramente las emociones y también que hay una correlación significativa entre prosodia y atributos pragmáticos.
Palabras clave: habla expresiva, clasificación, pragmática, análisis del discurso, emociones

Abstract: Although speech technologies keep improving their performance, it is necessary to understand the mechanisms used in speech to transmit, a part from lexical, other information such as emotion, attitude or speaker styles. In this work we have focused on the study of the correlation of basic prosodic features with emotional and pragmatic characteristics. For that purpose, three corpora have been used: emotional recorded speech, a read tale and Spanish parliament recordings. Based on an emotion and discourse modes taxonomy, we performed a classifying task of such characteristics by means of the prosodic features. Preliminary results show that emotions can be identified from prosody and that also exists a correlation between prosody and pragmatic attributes.
Keywords: expressive speech, classification, pragmatics, discourse analysis,emotion

1 *Introduction*

Current research community on speech technologies has an increasing interest on speech expressiveness. Most of the time, we have assumed that speech was emotionally neutral and that no other information than the words being said was transmitted in a human speech communication. Thus, the effort has been focused on *word* error rate and on neutral speech synthesis with just a suitable but meaningless prosody.

However, it is well known that there are other sources of information than lexical in speech and that human interaction is not only a matter of speech but also of common knowledge, gestures, non-verbal vocalisations, etc. This, for instance, has been understood by many people working on multimodal systems where information is gathered from different sources. Furthermore, although it is a hard task to compile, understand and model all these effects, there already are some research done on modelling emotion, as much in recognition as in synthesis (Nogueiras et al., 2001; Shröder, 2001).

In our work, we are focusing on information contained on prosody. It is well known that prosody can play several roles in speech. It is used to segment a discourse, to show emotional states, to focus the listener interest in a section of what is being said. It is also related to semantics and pragmatics (Prieto, 2003).

New applications such as speech-to-speech

* This work has been partially sponsored by the European Union under grant FP6-506738 (TC-STAR project, http://www.tc-star.org) and the Spanish Government under grant TIC2002-04447-C02 (ALIADO project, http://gps-tsc.upc.es/veu/aliado) and grant TIC2003-083820C05-03

translation, dialogue or multimodal systems demand for attitude and emotion modelling. In order to build more expressive and natural synthesizers we need to be able to model the effect of pragmatics on prosody. Humans would choose different ways to pronounce the same sentence depending on their intention, emotional state, etc. in a given context. The reasons that lead us to choose a specific prosodic configuration are rather pragmatic than lexical or syntactic. These claims support our interest on studying the behaviour of prosodic features related with a set of pragmatic attributes. Our work builds on the discourse modes definition made by (Calsamiglia and Tusón, 1999), we used her discourse modes as pragmatic attributes, and the idea of full-blown emotion discussed by (Cowie and Cornelius, 2003).

Here we present a first attempt to identify such events in speech. For that purpose we will try to classify emotions and pragmatics by means of prosodic features. We present three different frameworks: in one of them we will classify a set of four emotions, in the second one a set of discourse modes and in the last framework a subset of one of the previous modes.

For every framework we have collected a corpus, a set of sentences recorded by an actress, a tale story read by a professional and a parliamentary session. The pragmatic attributes were defined according to each corpus. Classification was carried out by means of a Multilayer Perceptron, a classical machine learning technique. We succeeded on classifying emotion labeled sentences. Therefore, encouraging results were obtained for the discourse mode classification of the tale. However, poor results were obtained while classifying discourse modes in the parliamentary database.

In the present paper we first describe the corpora and how they have been processed, then in Sections 3 and 4 we describe the feature set used and the emotion and pragmatic attributes we want to classify, in Section 5 we present the experiments done and finally in Section 6 we discus some relevant points of our work and in Section 7 conclusions that come up from the work are presented.

2 Corpora Description

The present work has been developed based on three different corpora. All of them were Spanish and each one recorded by one single speaker. Two of them recorded in a studio and another one recorded from a real situation. The first one of 1469 sentences was interpreted by a professional actress simulating four emotions, it was specially recorded for research purposes. We will refer to this corpus as `EMOT` from now on. The second one consists on a tale read by a professional. It is found data since it was recorded to be an audio-book and thus was not recorded for research purposes. The reader in this situation is free to read the text without any restriction. We will refer to it as `TALE` and it consists on 798 sentences. Finally, the third corpus was recorded from the Spanish parliament and it consists on a complete speech. It is composed by 220 sentences. It will be referred here as `PARL`. Each corpus has a specific recording conditions. While `EMOT` and `TALE` have been recorded in a studio, the `PARL` corpus has been recorded through satellite and is noisier.

All the corpus have been preprocessed automatically. Since `EMOT` and `TALE` were read we had the text. For the PARL corpus we had to modify the orthographic transcription published by the parliament in order to obtain a transcription that matched exactly what it had been said. Once we had an orthographic transcription of the text we applied the SAGA grapheme-to-phoneme system (J. Llisterri, 1993) in order to perform a phonetic transcription of all the text and the voice was segmented into phones by means of an HMM-based forced alignment (Adell et al., 2005).

Finally, the emotional and pragmatic labeling of the corpora was done manually by two people using the Transcriber tool (Barras et al., 2001).

3 Emotional and Pragmatic attributes

Each corpus requires a specific expressive analysis. Then, here we present the emotional an pragmatic classification of the sentences for every corpus.

3.1 Emotions

Here four basic emotions have been considered: *Neutral, Angry, Sad* and *Surprise*. We have chosen them since they are part of the widely considered full-blown or basic emotions group (Cowie and Cornelius, 2003).

3.2 Pragmatic

In the pragmatic discourse analysis performed by (Calsamiglia and Tusón, 1999), they present five discourse organization modes namely: *Narrative, Descriptive, Argumentative, Explanatory* and *Dialogue*. Each of these modes not only has specific characteristics in terms of lexical expressions, locutions but also distinctive prosodic dynamics. Therefore, they claim that each of these modes can be decomposed in several submodes as can be seen in Table 1.

Our work builds on this analysis, we used these discourse modes as pragmatic attributes in order to label the collected data. Each of the frameworks (`TALE` , `PARL`) has been given the set of attributes described below.

Mode	Sub-mode
Narrative	framework action resolution moral
Descriptive	physic place associative
Argumentative	framework facts conclusion
Explicative	-
Dialogue	-

Table 1: Discourse modes and their submodes. It is a summary of what is presented in (Calsamiglia and Tusón, 1999)

3.2.1 Discourse modes of a tale

When telling a story, events happen one after another, then the narrative mode is mainly present in it. However, it is necessary to describe characters, context, objects, etc. so descriptive mode is also present. Finally, sometimes characters speak directly themselves, this can be seen as a specific case of what Calsamiglia and Tusón define as the dialogue mode of the discourse. Thus, these three modes are the ones that have been considered to pragmatically label the recorded tale.

3.2.2 Discourse modes of Parliamentary speech

Although in a standard session of a parliament we could find examples of almost the five discourse modes, one of the most representative modes of the parliamentary discourse is the argumentative one. Thus, we have considered a parliament speech that was fully argumentative and it has been labeled using its sub-mode attributes. In this case the argumentative mode can be divided into *framework, facts* and *conclusion*. In fact, (Calsamiglia and Tusón, 1999) considers a wider set of sub-modes, however we have reduced them here.

4 Feature set definition

Here we intended to mainly analyze the acoustic features that carry prosodic information. One way to characterise pitch is by directly extracting parameters from the pitch contour. Another way we have used here to represent a F0 contour is by means of an intonation model parameters. We have choose the Fujisaki model (Fujisaki et al., 2000) together with the parameter extraction algorithm developed by (Agüero and Bonafonte, 2004).

The main acoustic parameters that describes prosody are *fundamental frequency, intensity* and *duration*. Thus, we have chosen a set of features that represent to some extend these parameters. Here it follows a short description of the acoustic features, grouped in the three mentioned classes.

Al the acoustic features have been extracted by means of PRAAT (Boersma and Weenink, 2005). The pitch contour was extracted every 5ms and smoothed using a low-pass filter with a 5Hz cut off frequency. Intensity contour was extracted in dBs and also every 5ms .

- **Intonation (direct measures):** We have considered some statistics measured directly on the F0 contour: *mean, standard deviation, maximum, minimum* and *range*. Also same features have been extracted from the *first* and *second derivatives* of the F0 contour. Additionally and due to well known effects on sentence beginnings and endings, the *mean F0 of the first and last voiced phone* are considered. Therefore, we calculate the mean, maximum, minimum and range of the *F0 phone means* and again these four measures only for *accented phones*.
- **Intonation (pitch modeling):** Fujisaki's model is based on phrase and

accent group commands; phrase commands are described by their position and amplitudes; accent commands by their position, amplitudes and durations. Here we are using the *mean value of the amplitudes*, the *mean value of the accent command duration* and the *number* of minor phrases and accent groups.

- **Duration:** This parameter can be understood also as rhythm. Rhythm is conformed by the duration of the units present in speech. We have considered a couple of units for that purpose: *syllables* and *phones*. Then, we calculated *speech rate* as number of syllables and number of phones per second. For the same reason as for intonation we incorporated the *duration of first and last syllables and phones* of the sentence. Since duration is known to have a log-normal distribution we also included the *logarithmic values* of the mentioned features. Silence also affect speech rhythm. Thus, additional features in this direction are the total *amount of silence* in the sentence and the *ratio total silence-total sentence time*.

- **Intensity:** Extracted features were the *mean, standard deviation, minimum* and *maximum* of the contour. Same values were calculated for the *first and second derivatives*.

To summarize, we are considering a set of 23 features directly extracted from the pitch contour, 5 features based on Fujisaki's model, 14 duration-related, 12 intensity-related features and 3 linguistic ones. A total set of 57 features are extracted from every sentence of the corpora.

5 Experiments

5.1 Description

In order to investigate the capabilities of the features to model the emotional and pragmatic attributes described here a classification task for the three frameworks presented here have been performed. For these purposes a standard machine learning classification algorithm such as Multilayer Perceptron (MLP) has been used. The software was Weka (Witten and Frank, 1999). The MLP has been trained with 66% of the corpus for each experiment and tested with the rest; 25% of the training set was used for validation.

A sentence classification task was performed. The `EMOT` corpus was already recorded sentence by sentence, so no other pre-processing was needed here. However, both `TALE` and `PARL` corpora were recorded at once, so we used punctuation symbols such as: ".;!?" for sentence splitting. The whole features set was extracted from each sentence. Detected silences have been considered as phrase breaks for the Fujisaki's model analysis.

5.2 Results

Here we present the results for the three experiments. We wanted to find out whether prosody can be used to identify emotional and pragmatic cues, thus we will look at classification performance measures such as confusion matrices or F-measure values. Furthermore, we are also interested on which features can better describe these phenomena. In order to discus it we will look at the information gain of the prosodic features used here.

Here it follows the definition of the statistical measures used for the analysis of the results:

- **F-measure**: It measures the general classification capabilities of the system and can be defined as follows:

$$F = \frac{2PR}{P+R} \tag{1}$$

where P stands for precision, R for Recall and they can be defined as:

$$Precision = \frac{n_i}{m_i} \tag{2}$$

$$Recall = \frac{n_i}{m'_i} \tag{3}$$

where n_i is the number of elements correctly classified as class i, m_i the number of elements of class i and m'_i the number of elements classified as class i.

- **Information Gain**: It measures a specific feature representative capabilities of a class and it is based on the entropy measure (H).

$$IG(C,A) = H(C) - H(C/A) \tag{4}$$

$$IG(C,A) = \sum_{ij} p(c_i,a_j) log_2(p(c_i,a_j)) \tag{5}$$

where C is a set of classes, A is a feature and $p(c_i, a_j)$ is the probability of belonging to class c_i and the feature A being $A = a_j$.

5.2.1 Emotions

Results for the emotion classification task are satisfactory since prosodic cues have allowed us to correctly classify 82% of the sentences. They are shown on Table 2 and are slightly better for Neutral and Sad speech and confusion is bigger within Anger and Surprise.

Classified as...					
Ang	*Neu*	*Sad*	*Sur*		**F-measure**
101	2	1	16	*Ang*	0.77
1	**100**	17	3	*Neu*	0.87
3	7	**107**	9	*Sad*	0.85
37	1	0	**94**	*Sur*	0.74

Table 2: Confusion matrix and F-measure values corresponding to the emotion classification task.

For this task, the main information is placed in F0 related features. In Table 3 the 10 features with highest information gain are shown and only one feature is not related to F0. Three first features stand out since their values are noticeably higher and they are related to F0 mean and maximum.

EMOT	
InfoGain	**Features**
0.7357	maximum of phone F0 mean
0.7065	F0 maximum
0.6499	F0 mean
0.4855	F0 minimum
0.485	first voiced phone F0 mean
0.4196	first derivative F0 mean
0.4097	F0 standard deviation
0.4041	last voiced phone F0 mean
0.4032	second derivative Intensity max.
0.4011	second derivative F0 mean

Table 3: Ten features with highest information gain for the `EMOT` task.

However, it must be noticed that features related to Fujisaki's intonation model are not as relevant as statistics directly extracted from the F0 contour.

5.2.2 Tale

For the pragmatic attributes classification we have not reached as successful results as for emotions. However, 64% of the sentences have been correctly classified. This shows that there exists a relation between prosodic features and the discourse modes described above.

Classified as...				
Des	*Nar*	*Dia*		**F-measure**
61	17	16	*Des*	0.62
25	**46**	8	*Nar*	0.60
18	13	**63**	*Dia*	0.71

Table 4: Confusion matrix and F-measure values corresponding to the tale pragmatic classification task

However, none of the features used here reach a suitable information gain, since all of them are lower than 0.15. On Table 5 the features with highest information gain are presented.

TALE	
InfoGain	**Features**
0.1498	F0 mean
0.1288	1st derivative F0 mean
0.1261	F0 minimum
0.1150	1st derivative F0 std. deviation
0.1093	F0 maximum
0.1076	mean phone-based speech rate
0.1072	mean accent command duration
0.1033	1st derivative F0 maximum
0.1017	F0 range
0.0937	last phone duration

Table 5: Ten features with highest information gain for the `TALE` task.

Despite the low values, it can be seen how all kind of features, related with F0, rhythm or Fujisaki's model, contribute to the classifications tasks. We would like to point out that the main important features are the ones related to Intonation but also Speech Rate and duration of the accent group command from the Fujisaki's model are relevant.

5.2.3 Parliamentary

For this task we got the worst results. Classification accuracy shows that prosodic features are not able to identify any of the submode attributes presented here.

Only 41% of the sentence have been correctly classified. Table 6 show results. Furthermore, information gain values were too low to be reported here.

Classified as...				
Con	*Fac*	*Fra*		**F-measure**
8	12	0	*Con*	0.33
13	**22**	0	*Fac*	0.55
8	11	**1**	*Fra*	0.10

Table 6: Confusion matrix and F-measure values corresponding to the parliament pragmatic classification task

6 Discussion

In the present work there are several assumptions that can be analyzed in order to deeply understand the problem faced in the present work: First, differences between the collected corpora. Since we wanted to have a variety of styles, we needed to collect corpora from very different sources and thus recording conditions may have influenced in the accuracy of the acoustic features. In fact, the best results have been achieved on best recording conditions.

On the other hand, we have chosen an elementary set of prosodic features. It might be interesting in the future to extend or replace the feature set by other features that can more accurately describe the intonation dynamics as in (D. Escudero-Mancebo and V.Cardeñoso-Payo, 2002), or even a finer description of duration effects on speech as the ones describe in (Barbosa and Bailly, 1994). For instance, modelling of local dynamics of prosody is needed.

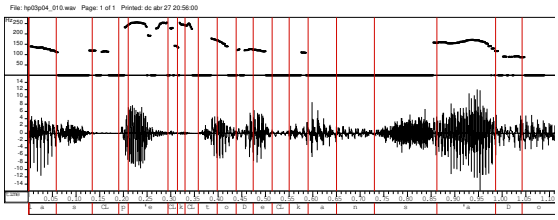

Figure 1: Pitch and phone segmentation for a piece of the sentence: *"Estaba sentado ante la mesa del desayuno, con aspecto de cansado y casi enfermo"*.

Usually emotion, style, discourse modes, etc. cause prosodic events only in a specific part of a sentence. For example, in Figure 1 there we have plotted the signal, pitch contour and phone segmentation of a piece of a descriptive sentence. It is very interesting to notice how the word *cansado* is emphasized by means of the lengthening of syllable `/sa/`. Since it is a descriptive sentence the main content word is this adjective and that is why it is strongly emphasized. The feature set used here cannot capture such local information since all its features are globally calculated all over the sentence shading all these effects.

Finally, speech characteristics other than prosody such as Voice Quality or formant movements may be correlated with pragmatic attributes. This means that there are other information sources that have not being exploited here and that might be successfully used for modelling emotional and pragmatic attributes.

7 Conclusions

Prosodic dynamics is affected by the speaker emotional state since we have been able to identify emotions by means of the prosodic cues. This has already been investigated but here we have been able to correctly classify 82% of the sentences.

For the pragmatic attributes classification problem on the `TALE` corpus, results have not been so encouraging but neither disheartening, since the classification task has been 64% succesful and that shows that prosody is used to mark discourse modes, what lead us to claim that prosody is influenced by discourse modes.

A harder problem has been to identify sub-mode attributes. We have encountered strong difficulties on classifying them.

It seems fairly clear that global prosodic features are useful but not enough to model the presented pragmatic attributes but results encourage us to work on the use of more local features for this task.

Poor results classifying discourse submodes does not mean that it is not possible to model them. We have shown how there exists a relationship between prosody and emotional-pragmatic attributes, but further research must be carried out on modelling this relation and maybe finding new features with more relevant information.

References

Adell, Jordi, Antonio Bonafonte, Jon Ander Gómez, and María José Castro. 2005. Comparative study of Automatic Phone Segmentation methods for TTS. In *Proceedings of ICASSP*, March. Philadelphie,PA,USA.

Agüero, Pablo Daniel, Klaus Wimmer and Antonio Bonafonte. 2004. Automatic

analysis and synthesis of fujisaki's intonation model for tts. In *Proc. of Speech Prosody*, March. Nara, Japan.

Barbosa, Plinio and Gerard Bailly. 1994. Characterisation of rhtytmic patterns for text-to-speech syntehsis. *Speech communication*, 15:127–137.

Barras, Claude, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 2001. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1–2). accepted for publication.

Boersma, Paul and David Weenink. 2005. Praat: doing phonetics by computer (version 4.3.04), March. http://www.praat.org/.

Calsamiglia, Helena and Amparo Tusón. 1999. *Las cosas del decir. Manual de análisis del discurso.* Ariel Lingüística, First edition. Chapter 10.

Cowie, Roddy and Randolph R. Cornelius. 2003. Describing the emotional states that are expressed in speech. *Speech Comnication*, 40:5–32.

D. Escudero-Mancebo, C. González-Farreras and V.Cardeñoso-Payo. 2002. Quantitative evaluation of relevant prosodic factors for text-to-speech synthesis. In *Porc. of ICSLP*.

Fujisaki, H., S. Ohno, , and S. Narusawa and. 2000. Physiological mechanisms and biomechanical modeling of fundamental frequency control for the common japanese and the standard chinese. In *Proceedings of the 5th Seminar on Speech Production*, pages 145–148. Bavaria, Germany.

J. Llisterri, José B. Mariño. 1993. Spanish adaptation of sampa and automatic phonetic transcription. Report SAM-A/UPC/001/V1, February.

Nogueiras, Albino, Asunción Moreno, Antonio Bonafonte, and José B. Mariño. 2001. Speech emotion recognition using hidden markov models. In *Proc. of Eurospeech*, September. Aalborg, Denmark.

Prieto, Pilar. 2003. *Teorías de la entonación.* Ariel Lingüística.

Shröder, Marc. 2001. Emotional Speech Synthesis: A Review. In *Proceedings of Eurospeech*, volume 1, pages 561–564, September. Aalborg, Denmark.

Witten, Ian H. and Eibe Frank. 1999. *Data Mining: Practical machine learning tools with Java implementations.* Morgan Kaufmann, October. http://www.cs.waikato.ac.nz/ml/.