Synthesized Speech : Naturalness, Subjectivity, Capture of Meaning

Geneviève Caelen-Haumont

SYNTHESIZED SPEECH: NATURALNESS, SUBJECTIVITY, CAPTURE OF MEANING

Geneviève Caelen-Haumont

Résumé

Il est maintenant couramment admis que les structures linguistiques ne peuvent pas rendre complètement compte de la forte variabilité observée dans le domaine de la parole. Cette variabilité est pourtant une composante essentielle de la communication. Il est donc nécessaire pour obtenir une parole de synthèse plus naturelle de modéliser cette variabilité. Les hypothèses basées sur une large expérience dans les secteurs de la lecture et du discours spontané, reposent sur l'idée que le locuteur pour donner à son discours les qualités du naturel, doit satisfaire à plusieurs conditions d'ordre pragmatique :

- Faire savoir le contenu du message, ce qui implique à la fois de le faire entendre (contraintes de démarcation et d'intelligibilité sur la forme linguistique), et d'autre part le faire comprendre en mettant en relief prosodique les unités linguistiques qui véhiculent l'information sémantique et/ou pragmatique (contrainte de discrimination) ;

- Faire croire le contenu du message. Pour susciter la croyance, puis l'adhésion voire l'action, le message doit comporter une dimension subjective, lieu de la rencontre intra-individuelle. Une grande part de cette dimension subjective réside dans l'excursion de la fréquence fondamentale (ou Fo) au sein d'un mot, et autres paramètres prosodiques qui lui sont attachés (durée, énergie).

Ces considérations s'appuient sur de nombreuses observations empiriques, et l'article présente un certain nombre d'exemples caractéristiques tirés de corpus de lecture et de corpus de parole spontanée, dans lesquels une forte amplitude est constatée au sein du mot lexical.

En synthèse de la parole, après la phase d'apprentissage de la structure linguistique, apprentissage de base qui est l'état actuel de la synthèse, une deuxième phase pourrait lui succéder. Ce serait celle où les systèmes s'affranchiraient localement d'une dépendance forte aux structures linguistiques normatives, afin d'adapter les formes à une expression plus subjective. C'est ce que font les enfants, bien que simultanément, dans l'apprentissage de leur langue maternelle.

Mots-clés : prosodie, synthèse de la parole, naturel, subjectivité, Fo, pitch range, sémantique, pragmatique.

Abstract

It is now well accepted that linguistic structures cannot completely account for the full variation that one observes in speech. This variation is nevertheless an essential component of communication. Therefore, in order to get more natural synthetic speech, it is necessary to model this variability. Based on experience in reading and spontaneous speech analysis, the grounding hypotheses of this work are :

CAELEN-HAUMONT, G. (2001), Synthesized Speech: Naturalness, Subjectivity, Capture of Meaning, *Travaux Interdisciplinaires du Laboratoire Parole et Langage,* vol. 20, p. 11-29.

1- the speaker needs to make the message known (both making it heard and understood),
2- in addition the speaker needs to make the message believed,
3- to be believed, a message has to supply a subjective dimension,
4- a great part of the subjective dimension lies in the Fo excursion within lexical items (and other related prosodic cues).
The justification of these claims is given in terms of empirical observations dealing with a number of examples of local variation in pitch range.
After the phase of linguistic structure learning (basic learning) which is the current focus of speech synthesis, another phase might be to break free with a strong dependency on normative linguistic links, in effect to adapt these forms to a more subjective expression. This is what infants do, albeit simultaneously, in their mother language learning.

Keywords : prosody, speech synthesis, naturalness, subjectivity, Fo, pitch range, semantics, pragmatics.

1. Introduction

The recent development of new fields of speech, sound, image, video technology and other domains using computer science, make users aware of and researchers sensitive to, the very important problem of naturalness. Among these different computer applications, speech is a very specific domain. A simplified image of a piece of reality, the outline of a portrait, a draft of an object may be fully satisfactory and even pleasing to see, and a sound needs not be sophisticated to be well accepted by the ear. However the ear identifies very well an artificial voice, and moreover it may not tolerate more than a few phrases of a rough or monotonous voice.

Naturalness is therefore a crucial challenge for speech technology. We need to invent or copy the conditions of naturalness, while there are adverse conditions to this naturalness, for instance the limited bandwidth of telephone channels.

2. Unreachable naturalness in speech processing : the phonetic level

Speech processing is a very difficult task in many respects. The so-called phonetic 'level' touches upon more fundamental human cognitive processes. This complexity is reflected in the related technological domains: recognition – understanding and synthesis.

Concerning the first domain, recognition and understanding, it is well known that any automatic processing is unfortunately unable to identify every phonetic segment. Several solutions can be used to achieve speech recognition. One of them is to make the system learn the lexical context (bigram or trigram) via a neural network and a great amount of speech data. This method may also lead to prosodic pattern learning [12]. Another way is to use expert rules in

order to enhance the system with another source of knowledge [8][9]. In any case prosodic information is used alternatively to provide a faster access to word information (lexical or grammatical, left and right boundaries), even before all syllables or phonemes have been fully identified. For more details, see a recent survey of the role of the prosodic information in automatic speech recognition and understanding [6].

None the less, whatever these approaches may be, they share a common feature: in this case, indeed, the phonetic stage, because of the problems encountered in the processing phase, tends to be circumvented.

In the synthesis domain, too, phonetics is a source of great difficulties. Although for several decades automatic processes and human expertise have made it possible to analyse more and more accurately the acoustic characteristics of phonemes, the problem of the inversion in synthesis is far from being solved: not only the natural and synthesized spectra are somehow different (for instance weaker amplitude of the formants, absence of speaker-dependent resonances), but a great deal of specific acoustic events characteristic of the human voice are lacking, such as disruptions, slight noises, unexpected voiced / unvoiced chunks ...

In fact, both in spontaneous speech and in reading, when the speaker is motivated to invest himself / herself in communication, a standard version of the phonemes never occurs. Speaker dependencies, rhythmic factors, linguistic context, acoustic and phonetic short-term and middle-term influences, attention paid to addressees, emotions ..., deeply alter the actual segment in comparison with the pure phonetic scheme; the latter only exists as model in our mind. This so-called phonetic variability is just the consequence of our lack of knowledge about all the processes involved in speech, where the linguistic component, though essential, is just an element among others. In the coming years progress is expected from further developments in the domain of articulatory models. For instance, modelling vocal tract, neuronal commands, organic interactivity and balance, sound propagation and modification in the vocal tract ..., might contribute to some extent to the improvement of synthesized speech. For a review of the challenges in the domain of synthesis, see [16] [17] [20].

Given this acoustic and phonetic complexity, the current trend in synthesis is close to the one in recognition and understanding, as mentioned above: circumventing the phonetic processing and substituting another method. In synthesis, the method used is the concatenation of small chunks of natural speech, generally overlapping two phonemes, — the so-called 'diphones'— , or more. Yet, the problem is not solved, it is bypassed.

Further studies are necessary regarding the cues of interactions at different levels and the role of emotion underlying any utterance with its prosodic correlates, notably at the lexical level.

This requires investigating the function of psychological investment in speech, in other words the personal, i.e. social and individual characteristics of speech. Still we should keep in mind that speech is not made of distinct layers but works as an integrated whole. This might imply a change in conceiving, recording and analysing corpora. Whereas the current tendency would be to cumulate layers of information from the acoustic level to the upper ones, a new approach would proceed from a top-down analysis perspective: from pragmatic and psycholinguistic conditions to the linguistic (semantic, syntactic) component, and ultimately to the phonetic and acoustic segments. To use a metaphor, in speech the segment is not 'coloured' superficially but in its depth.

3. About some characteristics of naturalness in speech

If at present we cannot claim to reach, hence predict, the characteristics of phonetic naturalness in speech synthesis, that is to say the acoustic variability of a phoneme (and its phases) in a specific context, still we can describe some of the characteristics of naturalness from a general point of view. These characteristics should be taken into account in synthesis.

When various fragments of spontaneous speech (and to some extent 'intelligent' reading as well) are submitted to analysis, a set of characteristics may be found, in short, variability, adaptability to communication context and addressees, and ultimately the subjective capture of speech characteristics at every level, from acoustics to semantics. The permanence and redundancy of linguistic structures, on the one hand, and the strength of the situation which greatly contributes to reducing ambiguity, on the other hand, give the speaker a relative freedom to disrupt this linguistic framework. In fact the phonemes are far from realising their canonical forms, various disfluencies break the 'right' (i.e. textual or academic) linguistic structure, and lexical prosody often disrupts the syntactic organisation. In spite of this, spontaneous speakers understand each other well, and often better than in the conventional speech of readers. Since the language model and structures may (or may not) be activated independently from the effective realisation of speech, speakers can 'appropriate' language forms at the acoustical, phonetic, prosodic, semantic, syntactic and/or emotional levels.

In dialogue conditions, it is observed that the form of speech is conditioned by the feedback about understanding or agreement that the speaker expects from the listener [22]. In the new exploration of this domain, some studies in prosody show that all these means of omissions, substitutions, repetitions, breaking and pauses, various noises and non-sense utterances, wide pitch excursions, supposedly disrupting the linguistic framework, lead on the contrary to better communication and interaction between speakers [10] [11] [15], as they provide cues of

synchronisation between speakers, and perhaps also facilitate automatic recognition and understanding [12].

4. How to prosodically converge towards more naturalness in speech synthesis

In my opinion, considering the great amount of research done and underway in the field of prosody, the analysis should not only deal with linguistics but also with pragmatics taking into account all the elements of speech situation and conditions, and in this domain, the speaker's (i.e. subjective) point of view is prevalent. More explicitly, each domain, linguistics and pragmatics, may claim to integrate the scope of the other one, as the foreground facts of a domain are also background or sideground facts of the other one. What remains unexplained within a domain is treated as variability and accounted as statistical variance. Conversely, from the viewpoint of another paradigm, it may be taken into account as a significant aspect of speech reality. For instance, spontaneous speech syntax, a typical issue in pragmatics, could be envisaged from a normative perspective, a gibberish of wrong forms or mistakes, and thus rejected. In addition, if the social motivations of speakers may be considered as driving the organisation of speech, subjective instances of speech may then remain unnoticed or be discarded. Many more examples of this discrepancy between perspectives could be proposed.

Thus, a crucial question arises in the domain of speech synthesis: if these perspectives capture only certain aspects of reality, often with contradictory basic assumptions, which model of speech do we have to apply? Moreover, current synthesis models are almost exclusively borrowed from linguistics. A reading model for dialogue purpose in real life is inappropriate, although it might be adequate for instance for displaying technical documents or contents of e-mails messages to blind users.

The proper way is probably to be as close as possible to the conditions and circumstances of speech. Both in human and automated speech analysis, pragmatics is certainly a better paradigm as it is more general. If we take it for granted that pragmatics encompasses a linguistic perspective, evidently in certain conditions of communication, or in particular moments of speech, pragmatic requirements would be nothing else than pure linguistic constraints. This approach leads to the idea that the main perspective in prosodic analysis might be essentially oriented towards the speaker's point of view. It might be the link between linguistics and pragmatics and help unifying these different perspectives. Natural communication is a matter of person-to-person relation, not a relation between conceptual systems, and in this relation men and women have at their disposal a great deal of resources and tools, which include the linguistic ones, of course, but also para- and extralinguistic ones such as pause, prosodic effects,

disruptions, disfluencies... If we really intend to come as close as possible to a natural speech expression, we need to encapsulate this kind of characteristics in speech synthesis.
Indeed, the choice of words, phrase ordering, sentence structures, i.e. the semantic and syntactic means, contribute in framing and casting the meaning in the most appropriate way. Still all the paralinguistic and extralinguistic stuff is superimposed to clarify, clearly disambiguate, capture meaning in a subtle, personal way. This stuff is the matter of shared codes; however its use, occurrence and combination in the actual performance stand for an accurate and personal capture of sense.
Thus, this capture outlines a sort of subjective space whereby the only way to subjectively express meaning is to prosodically modify, release, or set *against* the well-framed organisation of linguistic units: for instance, using unexpected prominence with respect to the syntactic status of the word, or opposing a prosodic grouping (and/or pause) to the syntactic one... Evidently — except in the case of very grave speech disorders — the linguistic organisation may never be broken in practice, because it is a social convention and therefore a reality independent from its actual realisation, which stays apart from the prosodic outputs. This gives a measure of the relative great freedom given to each speaker to prosodically modify (i.e. capture) the links between linguistic forms (and to some extent, contents) in speech. Even though this linguistic organisation may not be a straightjacket, as the speaker is free to choose lexical items, contexts and combinations, it remains a social convention, something still external and somehow impersonal [3] [5]. In fact this impersonality reflects the present situation of our synthetic speech outputs.

5. Linguistic structures, prosody and the capture of meaning

In our experience of reading and spontaneous speech in French, prosody encompasses two manners of expressing the meaning: in addition to the intonation line which conveys a linguistic framework (essentially the grouping function), the melodic excursion in the local (lexical) domain conveys a subjective relation to the meaning. In this relation, the speaker's freedom consists in attributing relative prosodic prominence to lexical items, notably Fo prominence in terms of maximum and range. In brief, the more the Fo line in the lexical unit is deviating from the mean Fo line of the whole prosodic group (i.e. intonation), the more this lexical unit (and therefore the group) expresses a subjective capture of meaning and/or the speaker's communication intention [3] [4]. This Fo local excursion in the word is the place of the speaker's discrepancy, *or not,* vis-à-vis the canonical prosodic expression of linguistic structures. It defines a space of relative freedom at the level of words in the prosodic group.

Of course these two processes may be combined when, in French for instance, such a word coincides with the right boundary of a group. In my opinion, combined or not, these two processes are distinct. If the prosodic expression of linguistic structures has prevailed in synthesis until recently, the consideration of the subjective capture of meaning would significantly contribute to personalizing the artificial speech outputs, thereby improving naturalness.

Let us give more precision to this analysis. In French and in other languages, prosody plays two main functions to convey meaning. First it expresses well-known linguistic functions such as syntactic and semantic ones, both structural (for instance in the semantic domain, theme-rheme organisation). These functions belong to the domain of intonation (Fo, but also timing and intensity) shaping sentences and phrases. At this stage the speaker does not invest him/herself in prosodically reformulating the linguistic links between units (and therefore specifying a subjective meaning). The speaker only gives way to their own linguistic competence. This conventional prosody may however fit with their voluntary or involuntary purpose.
Based on other considerations than syntactic, phonologic or semantic ones, another grouping may be used. It is based on rhythmic patterns extracted from readings in psycholinguistic experiments. These word groupings extracted from readings in psychological experiments concerning performance structures [13], when applied to synthetic speech [24], result in more natural, more fluent speech, presumably because they correspond to actual encoding units [3] [24]. Nevertheless, as mentioned in Zellner's paper, this improvement concerns a neutral reading which does not convey particular emotional, semantic or pragmatic content. Therefore it may be considered an initial but significant step towards more naturalness.
Secondly, prosody expresses subjective meaning, whose domain is local. It concerns the lexical organisation of melody and related other prosodic parameters already mentioned.
As in other fields, in speech a person settles their identity by discarding common behaviour to some extent. A space remains free for each speaker, *given the linguistic rules and intonative background,* to disrupt and break down, (-or conversely to support and even to focus) the syntactic links between units [2][24]. More precisely, this space is prosodically outlined by the Fo range within words (in fact |ΔFo| because in this space the relevant information lies in the difference between Fo maximum / minimum, not in the direction of the Fo slope), and associated cues such as Fo maximum, duration, and occasionally, intensity, pause, downstepping.
Thus, untying the linguistic link between items is a way of expressing subjective meaning — which might be what 'meaning' stands for. This process makes sense because the actual linguistic structure and intonation are interpreted in terms of the speaker's / listener's linguistic

and prosodic knowledge. This intimate use of meaning is a sort of a game playing around linguistic units; none the less an essential process for the complete understanding.
In fact, prosody is a sort of trade-off between two antagonistic forces, on the one hand a trend towards social convention, structure, norm, an external point of view, and on the other hand, towards subjective expression, local viewpoints, rupture, identity, emotion, present realisation. The first one provides cohesive strength, and in the same time, the second one tends to disrupt this continuity (dissociative strength).
The first one refers to the social norms of linguistics and its prosodic counterpart, intonation, which both cannot be altered without compromising the understanding. The second one corresponds to the subjective use of the linguistic structures, which works at the local / lexical layer. While speakers resort to social norms to prosodically and locally recompose the linguistic links between words at their convenience, the choice of the items to be more or less prosodically focussed, even involuntary, is their own. It is a space of subjective freedom, a way of capturing meaning.
This capture of meaning is all the more important because in spontaneous speech, for instance, background information is not supplied as it would be in a text or prepared speech. This lack of conceptual accuracy is compensated by the paralinguistic and extralinguistic precision delivered all at once in speech, in which prosody plays an essential role.
In reading, the speaker is not the author. Thus the only possibility to invest him/herself in the conceptual domain of the text is to set up a distance vis-à-vis linguistic structures, a space of one's own, in which everything is checked and recast in function of personal interpretation and feeling. In other terms, when a reader, or more generally a speaker, is talking and conveying their own point of view, as said before, they can only deliver their own feeling against the background of linguistic structures, simply because the language system allows this personal capture. This instantaneous, actual and short-lived speaker's filtering of linguistic meaning is an essential prosodic function. In the second part of the paper, examples will support this hypothesis.

The next development concerns the melodic excursion within words, from a local and subjective perspective.
Figure 1 below is a fragment of reading in French. In this chunk, for instance, the speaker's prosody obviously recasts the syntactic structure. If the right boundary of the NP1 ('*vers*') is highlighted, nevertheless the F0 range (|ΔF0|) is smaller than the one of the right boundary ('*marin*') of the prepositional NP, which, moreover, is syntactically of a minor level and dependent (i.e. embedded).

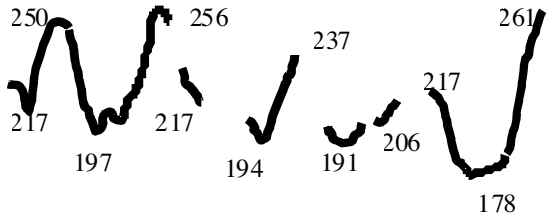

Figure 1
Male speaker.
Fragment of reading in French extracted from the sentence: "Ces longs vers prospèrent sur le plancher marin des zones sous-marines profondes." ("These long worms are prospering in the deep areas of sea bed"). *The numbers correspond to the minimum and maximum Fo values in Hz.*

In the same fragment, due to the semantic field in progress ('giants worms' isotopy which is the main theme of the text), a wide range is given to the lexical word *'longs'*. The widest one is attributed to the word 'marin' which is the first occurrence of an unexpected information (i.e. a very deep sea bed is a hospitable place for worms). This chunk of speech displays a relevant example of linguistic structures captured and linguistic links reshaped: pragmatic (and semantic) considerations such as for instance taking into account their addressees, are in the foreground, and prosody enables to highlight this process.
In my own perspective, this play (or in other words, this 'dialogue') between, on the one hand, subjectivity, and on the other hand, linguistic structure, is the closest way of describing the real nature of speech, that is properly, its subjective and effective dimensions.

6. Main prosodic functions in speech generation

I would like to bring to light, or define more precisely, some prosodic functions that seem to be important in the context of communication and understanding.
In my opinion, as already expressed, the main function of prosody that governs the other ones, is a pragmatic one, and subjective expression is a window opened in this field. In such a perspective, prosody may be viewed as playing two main functions.

6.1. Making known

One of them is *'making known'*. In this domain, the goals of speakers and generation systems are twofold: on the one hand, making the linguistic units *well heard,* that is to say, allowing a

proper demarcation at the phonetic/word level (for instance, in French prosody, pitch and timing are crucial in left and especially right boundaries), but also at the group and sentence levels, which is properly the role of intonation. This well-known prosodic sub-function supports linguistic structures, but also probably, carries to some extent, *linguistic* information (such as lexical vs. grammatical syllable, beginning / end of a lexical word), which facilitates the human decoding process [8][9]. In the 'making heard' domain, prosody is mainly dealing with the whole linguistic form. Here the pragmatic function of prosody fits best with the linguistic one.

On the other hand, the second sub-function is to make utterances *well understood*. This goal requires particular focus on specific units from the overall stream, the very ones that seem to carry the main information (whatever it could be) from the speaker's point of view. The following illustrates these considerations.

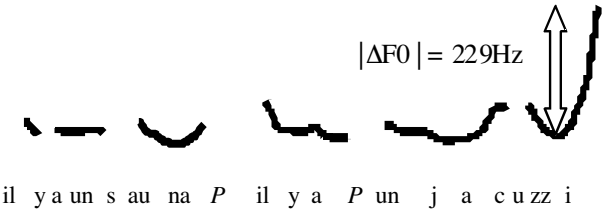

Figure 2

Female speaker.

Fragment of spontaneous speech: "... il y a un sauna, il y a un jacuzzi ..."
(... "there is a sauna, there is a jacuzzi" ...). |ΔFo| *is expressed in Hz. P corresponds to a pause.*

Figure 2 above displays an example of a word which is not common *('jacuzzi')*. This word belongs to a syntactic group (the third one from a sequence of six, each of which has a noun in final position). Interestingly, the range of this word is the greatest (229 Hz) from the whole sequence of these six syntactic groups.

According to the needs of their own expression or the estimated needs of their addressees, the speaker then adjusts the pitch range to the relevance of words in the current pragmatic and subjective conceptual model.

Figure 3
Male speaker.
Fragment of reading: "Des sources thermales chaudes y maintiennent une température moyenne élevée." ("Hot springs keep a high mean temperature").
The pitch range (|ΔF0|) *is calculated in 1/8th tones.*

In the example of figure 3 above, the word *'sources'* ('springs') is given the biggest pitch range as it is unexpected in the context of deep sea beds. The end of the NP1 *'chaudes'* ('hot') is not highlighted at all: comparatively, pitch range is rather flat (4 eights of a tone), and there is no pause after it.

The next example below (figure 4) presents the same phenomenon of unexpected information in spontaneous dialogue conditions, although in a specific situation and context of speech: a *tropical* plant greenhouse in French mountains area. The word *"tropical"* displays the widest pitch range (|ΔF0| = 120Hz) in the whole sequence.

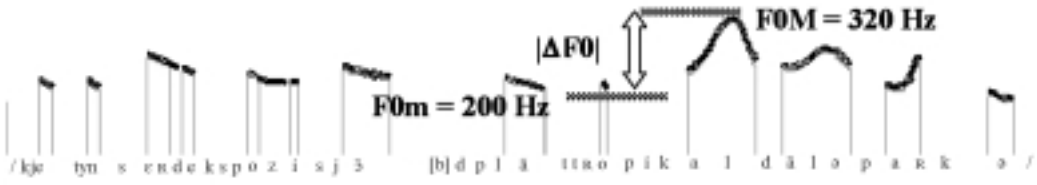

Figure 4
Female speaker.
Fragment of spontaneous speech in dialogue conditions : «... qui est une serre d'exposition de plantes tropicales dans le parc » *("which is a greenhouse of tropical plants exhibition in the park").*
|ΔF0| *is expressed in Hz.*

This prosodic function is given some more concrete evidence, such as for instance in French, when a lexical item which is highlighted does not coincide with the right boundary of a group which is usually accentuated. Indeed, both prosodic events may come together at this place.

6.2. Making believed

It remains that in speech communication, 'making known' information is not enough, as a main dimension is the expression of beliefs. Thus, an important prosodic function is *making believed*. Though 'making believed' and 'making known' refer to two different functions, nevertheless they cannot be isolated in the prosodic process. The only way to be believed is to make known which lexical units are conveying at the best our belief and personal truth[1]. Prosody needs to be convincing of its own on top of the linguistic structure. *The more the speaker invests her/himself in speech, the more they try to be convincing and the more they do so by the way of prosody, thereby evading from the regular linguistic framework.*

If prosody is just used as a sort of acoustic paraphrase of linguistic structures, of course the meaning may be available, but it is ill-instantiated in actual speech conditions, and no information is supplied to guide its interpretation. In this situation, the speaker cannot — or will not— deliver a personal interpretation of the linguistic structures. Sometimes this prosodic expression may lead to a better understanding, when the listener acknowledges the speaker's intent and prosodic compliance with speech conditions. This is the case when the speaker refers to an external authority's text or discourse. Here the listener may not expect a subjective meaning conveyed by prosody. Anyway, even though this kind of style might sound right according to the situation, it becomes rapidly unpleasant and boring.

In fact, this effect of prosodic weariness seems to be reached not only because of the repetition of the same syntactic patterns, but also because the person is perceived to some extent as 'absent-minded' in their speech. For instance, belief, a component of a motivated speech, and — consciously or not — perceived as a strong expression of the speaker's personality, is entrusted to another person or authority in the weary style of speaking. So, pragmatic conditions (hic, nunc, ego) are not exactly performed, and the interest is not aroused.

Moreover, interest in speech is aroused when a person's belief is conveyed and when some kind of innovation takes place. Prosody, in fact, has to say more than syntactic structures can do. Further it brings to the foreground unpredictable meaning at the very moment the listener is decoding utterances, by focussing or lowering the importance of words. This process represents at the same time the condition of linking subjectivity to speech and improving understanding by listener(s), as subjectivity (feelings, beliefs) is made accessible and offered to be shared. Beyond linguistics, a communication process is at work between two persons who recognize each other because they share the same psychoprosodic use and the same rules.

1.Therefore, all the examples presented above and below in this paper illustrate simultaneously these two functions.

To be more precise, natural speech, when motivated, is never departing from a kind of 'emotional' expression. The focus is on "ordinary" emotion underlying human communication, in the absence of strong ones. This emotional expression is often under mind control, but sometimes it is not, of course. According to my experience, in both situations, the differences between melodic ranges in words are mainly filtered by this emotional component, as the speaker is expressing a feeling, a personal point of view superimposed on the linguistic stream. Speaker identity and subjectivity prevail, they stand in the foreground. This perspective fits best with other studies in the field of prosody and emotion [25].

This means that prosody supplies implicit meanings superimposed to the linguistic meaning (and the activation of associated semantic networks).

First, referring to the conditions and context of speech, an implicit meaning could be translated in terms such that: "here this word expresses my feeling that...". The contents of this feeling might be such as for instance: "no doubt that I'm right", or "mind, you don't expect this word", or "just consider this word, it will be important later", or otherwise, "don't take care of this one, it has no relevant meaning, it is simply a bridge to the next one"...

Secondly, beyond this function of conveying the expression of one's own feelings or taking care (or not) of the addressees' ones, prosody may also express other implicit meanings, in the case of attitudes, for instance irony, and especially in dialogue conditions, in the case of indirect speech act.

Figure 5 below displays an example of irony. In the previous sentence, the speaker was mentioning a street previously called *'rue de Lyon'*. The prosodic mechanism of irony works at two levels: first, the word *'simplifié'* gets a wide range (250 Hz), secondly, the following sequence is clearly lowered. We notice that even the informative part, i.e. the new name of the street *('Hiskovitch'),* is not only focussed but lowered too. This is another clear illustration of the distribution of roles between linguistic structures, which convey information, and lexical prosody which puts an attitude in the foreground.

Figure 5
Female speaker.
Example of irony. Fragment of spontaneous speech: « on a simplifié et maintenant elle s'appelle la rue Hiskovitch » ([the name] has been simplified and now it is called the street Hiskovitch).
|ΔFo|| *is expressed in Hz.*

Figure 6 below is the next sequence of the same sentence.

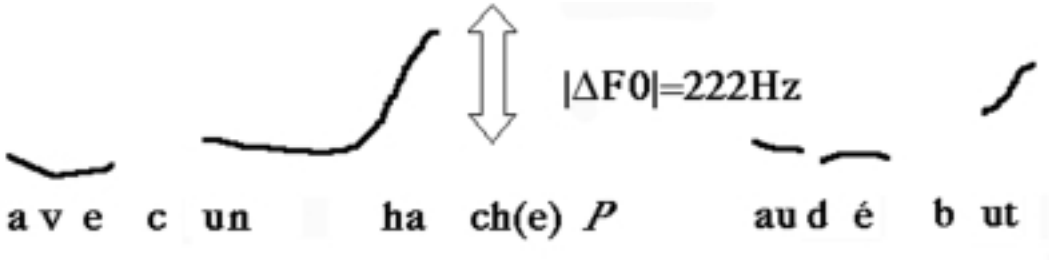

Figure 6
Same speaker, same corpus, same sentence ... "avec un h au début" ("with an h at the beginning").
|ΔFo| *is expressed in Hz.*

In this example, the informative part of the sentence (phonetically *"hache",* i.e. 'h'), which expresses a metalinguistic purpose clarifying the spelling of the name *'Hiskovitch',* is again associated with a wide pitch excursion (|ΔFo|=222Hz) as expected, and a pause (P). This melodic range is wider than that of the syntactic group boundary *'début'* which is nevertheless hierarchically higher and independent.
In the area of indirect speech act in dialogue, within a given linguistic context (for instance: /it is hot here/), prosody makes it possible to identify an illocutionary act as a question or a statement, and for instance to prompt the listener to act (for instance, here, to open the window). Thus, in the case of irony or indirect speech act, for instance, prosody alone, possibly, or with the support of situation, may convey meaning beyond linguistic items, and even, in their place.

In such a function, prosody works precisely as a paraphrase or an antiphrase.

7. Towards some improvements in speech synthesis naturalness

As already mentioned before, some characteristics of naturalness in speech may be expressed in terms of variability, diversity, spontaneity, adaptability and subjective capture. The consideration of these properties leads to some recommendations to improve synthetic speech. In correspondence with our preceding remarks, these directives deal with three dimensions of speech: context situation, linguistic and subjective levels. Only a few of them will be exemplified relative to these levels.

7.1. Linguistics and diversity

For any language, other networks of significance in the linguistic frame need to be explored and applied even in the syntactic domain. For instance, among others, dependency models [1][3][4], first originated from Tesnière's theory [21], may be adapted and generalised. Theme / rheme organization, especially viewed from a hierarchic and generative perspective [3][4][23] is also convenient, and should be checked. Because they allow to prosodically express different perspectives on linguistic stream and therefore linguistic meaning, these different linguistic models are relevant, and furthermore, may supply more diversity. This kind of prosody concerns linguistic structure. It may be used alternatively as such, or combined with subjective models that recast the linguistic framework.

7.2. Speech context and adaptability

Adaptability to the speech context may reflect different kinds of sociological needs. If synthesized speech is profiled to be relevant, it has to suit users' demands, such as for instance regional languages. Another field of social application is to provide different styles. In reading, concurrently to prosodic structural models mentioned above, and dictation, a prosodic style based on social grounds could perfectly fit. For instance, following a first processing in the structural domain, another processing based on the semantic network adapted to the situation might be undertaken. This semantic network might also be filtered in compliance with subgoals of the communication to fulfil the needs of a certain kind of addressees. Arriving at a spontaneous-like speech will indeed require a great amount of additional research. First of all, wider pitch excursions, different kind of fillers, more prosodic disruptions and contrasts displayed in the structural models might be nevertheless checked as they might more directly contribute to a natural expression.

7.3. Subjectivity and the capture of meaning

As explained before, a standardised speech seems to stand far from the natural one, insofar of course as all the ingredients composing a concrete situation of communication tend to discard standardisation. Thus, in my own perspective, in synthesis, a model based on average structures or average speakers leading to standardisation, is irrelevant if it is not reworked. Models need to encompass singularity, which is characteristic of natural utterances. Singularity, in turn, may be reached in the subjective space of prosody at the local / lexical level. An interesting challenge is to try to reproduce this inner prosodic trade-off between linguistic structure and subjective expression, which is the private game of taking a distance from lexical items or of appropriating their meanings. A way of approaching this intimacy is to be carefully sensitive to the speaker's subordinate and superordinate goals and feelings. Goals and feelings are one of the main roads that lead to subjective expression. They are also effective in constructing a classification between lexical items for the specific needs of speech synthesis.

Another way of expressing prosodic subjectivity is to alternate linguistic and / or subjective models. According to what has been observed in spontaneous dialogue [5] and intelligent reading [3][4][8], speakers base prosodic expression, and especially pitch range, on the underlying linguistic (syntactic or semantic) or subjective (feeling of complexity of the word contents, lexical field continuity, or unexpected information...) structures or networks which they are sensitive to at this present moment of their speech.

These results are in line with other studies in psycholinguistics, based on semantic purposes more than on syntactic ones [18][19], for instance the aspects of 'transitory understanding', and with the idea of competitiveness between several fields of information in speech [14].

In most cases, and within the limits of this experiment [3][4], the linguistic and subjective models provide relevant pitch ranges covering a few phrases –i.e. some minimal prosodic groups– regardless of the sentence frame; as already mentioned above, the unit of conceptualisation is based on prosodic groups and not on sentences. In these experiments, the connection between the successive models at phrase boundaries was important and it was observed that the melodic transition from one to the next one was always smooth. As two successive and different underlying models connected at group boundaries are agreeing with each other, a hypothesis may be put forward that the final tone of the preceding group (and model) may act as a priming for the next one (and model). Both in production and in understanding processes, this facilitates the transition from an underlying model to the next. Anyway, for the purpose of speech generation, the melodic transition between two different underlying models needs to be smooth.

8. Conclusion

To summarise, prosody plays a linguistic function when it highlights phonetic, morphosyntactic, syntactic or semantic *structure.* In this case, the pragmatic function of prosody is restricted to the linguistic one. The reference to speaker subjectivity may then be minimal. This style of prosody may be useful when the speaker cannot — or does not want to — invest or express his feeling, but it is insufficient, or even irrelevant, when speech is subjectively motivated.

In this case another prosodic line is woven into the lexical dimension of the linguistic stream superimposed to the intonation baseline, and the Fo range (|ΔFo|) is assigned the main role. By this very fact, this prosodic style enhances the expression of belief, and it is greatly subjective. It contains the prosodic signals (and impulses) for giving rise, among addressees, to interaction.

This paper aimed at improving naturalness in speech generation and synthesis by reintroducing the speaker's point of view on linguistic structures, and superimposed to them. This consideration leads to the idea that once the syntactic level is prosodically settled, another process could be undertaken, modifying local lexical pitch range, places and levels of some Fo maxima values, pause and duration according to users' goals. This process can be grounded on the lexical domain by taking into account semantic and pragmatic considerations during the generation phase.

In my perspective, the new challenge for speech synthesis may be expressed as such: how to give an accurate, live, convenient, or personal meaning which reconfigures the links between linguistic units *without losing anything* from the linguistic meaning, and without being artificial.

9. Acknowledgements

Kind thanks to B. Bel for the help given in the English expression.

References

[1] G. BAILLY, "Integration of Rhythmic and Syntactic Constraints in a Model of Generation of French Prosody", *Speech Com.*, 1989, p. 137-146.

[2] G. CAELEN-HAUMONT, *Structures prosodiques de la phrase énonciative simple et étendue*, Ed. J.-P. Köster, Hamburger Phonetische Beïtrage, Band 34, Hamburger Buske, 1981.

[3] G. CAELEN-HAUMONT, *"Stratégies des locuteurs en réponse à des consignes de lecture d'un texte : analyse des interactions entre modèles syntaxiques, sémantiques, pragmatique et paramètres prosodiques"*, Thèse de doctorat d'état, Université de Provence, Aix-Marseille 1, vol. 1 et 2, 1991.

[4] G. CAELEN-HAUMONT, "Synthesis: Semantic and Pragmatic Predictions of Prosodic Structure", chapter 13, *Fundamentals of Speech Synthesis and Speech Recognition*, Ed. E. Keller, J. Wiley and Sons, Ltd, Chichester, England, 1994, p. 271-293.

[5] G. CAELEN-HAUMONT, "Du faire-savoir au faire-croire : aspects de la diversité prosodique", *Traitement Automatique des Langues*, ATALA-CNRS, Paris, vol. 38, n°1, 1997, p. 5-26.

[6] G. CAELEN-HAUMONT Guest ed., *CC-AI Special issue "On the role of Prosody in Automatic Speech Recognition and Understanding"*, 15, n° 3, 1998, p. 151-305.

[7] G. CAELEN-HAUMONT, B. BEL, "Le caractère spontané dans la parole et le chant improvisés : de la structure intonative au mélisme, *Parole,* 1, vol. 15, 2001.

[8] B. CAILLAUD, *Apprentissage de connaissances prosodiques pous la reconnaissance automatique de la parole*, Thèse INPG, Grenoble, France, 1996.

[9] B. CAILLAUD, P. MUNTEANU, J.-F. SERIGNAT, J. CAELEN, "Knowledge Acquisition for Lexical Access Improvement", *CC-AI Special issue "On the role of Prosody in Automatic Speech Recognition and Understanding "*, Guest-Ed. G. Caelen-Haumont, 15, n° 3, 1998, p. 255-278.

[10] H.H. CLARK, "Speaking in Time", *Proc. of ESCA International Workshop on Dialogue and Prosody,* De Koningshof, Veldhoven, The Netherlands, 1999, p. 1-6.

[11] K. FISCHER, "Discourse Effects on the Prosodic Properties of Repetitions in Human-Computer Interaction", *Proc. of ESCA International Workshop on Dialogue and Prosody,* De Koningshof, Veldhoven, The Netherlands, 1999, p. 123-128.

[12] F. GALLWITZ, H. NIEMANN, E. NÖTH, V. WARNKE, "Prosodic Information for Integrated Word-and-Boundary Recognition", *Proc. of ESCA International Workshop on Dialogue and Prosody,* De Koningshof, Veldhoven, The Netherlands, 1999, p. 163-168.

[13] J.P. GEE, F. GROSJEAN, Performance structures: A psycholinguistic and linguistic appraisal, *Cognitive Psychology*, 15, 1983, p. 411-458.

[14] M. HUPET, J. COSTERMANS, "Et que ferons-nous du contexte pragmatique de l'énonciation", *Bull. de Psychologie,* XXXV, n°356, 1981-2, p. 759-766.

[15] S. JEKAT, "Prosodic Cues as Basis for Restructuring", *Proc. of ESCA International Workshop on Dialogue and Prosody,* De Koningshof, Veldhoven, The Netherlands, 1999, p. 135-137.

[16] E. KELLER and B. ZELLNER (eds), *Les défis actuels en synthèse de la parole,* Etudes de Lettres, Université de Lausanne, Suisse, 1997.

[17] E. KELLER, "Les théories de la parole dans l'éprouvette de la synthèse", *Les défis actuels en synthèse de la parole,* Eds. E. Keller and B. Zellner, Etudes de Lettres, Université de Lausanne, Suisse, 1997, p. 9-27.

[18] W. KINTSCH, T.A. VAN DIJK, "Toward a Model of Discourse Comprehension and Production", *Psychological Review,* 85, 1978, p. 363-394.

[19] J.-F. LE NY, M. CARFANTAN, J.-C. VERSTIGGEL, "Accessibilité en mémoire de travail et rôle d'un retraitement lors de la compréhension de phrases", *Bull. de Psychologie,* XXXV, n°356, 1981-2, p. 627-634.

[20] J. LOCAL, "Ce qu'on peut faire pour la synthèse de la parole avec un peu plus de prosodie et une meilleure qualité de signal", *Les défis actuels en synthèse de la parole,* Eds. E. Keller and B. Zellner, Etudes de Lettres, Université de Lausanne, Suisse, 1997, p. 29-46.

[21] L. TESNIÈRE, *Eléments de syntaxe structurale,* Ed. Klincksieck, 1959, 1965.

[22] R.S. TOMLIN, L. FORREST, M.M. PU, M.H. KIM, "Discourse Semantics", chapter 3

in *Discourse as Structure and Process, Discourses Studies: a Multidisciplinary Introduction,* Vol. 1, Ed. T.A. Van Dijk, Sage Publications, London, 1997, p. 63-156.

[23] C. TOURATIER, "Structure informative et structure syntaxique", *BSL,* LXXXVIII, fascicule 1, 1993, p. 49-63.

[24] B. ZELLNER, "La fluidité en synthèse de la parole", *Les défis actuels en synthèse de la parole,* Eds. E. Keller and B. Zellner, Etudes de Lettres, Université de Lausanne, Suisse, 1997, p. 47-78.

[25] B. ZEI, "Au commencement était le cri", *Le Temps Stratégique,* 1995, p. 96-103.