Articulation Degree as a Prosodic Dimension of Expressive Speech

Grégory Beller, Nicolas Obin and Xavier rodet

IRCAM - Institut de Recherche et Coordination Acoustique Musique
1 place Igor Stravinsky, 75004 Paris
beller@ircam.fr nobin@ircam.fr rodet@ircam.fr

Abstract

In this paper, we present a study on the influence of the expressivity on the articulation degree. It is part of a project that aims to transform the expressivity of an utterance for artistic purposes, such as cinema, theater and contemporary music. An algorithm for formant frequency estimation is presented. Then a measurement of the articulation degree is proposed involving a joint statistical analysis of the vocalic triangle area and of the speech rate. The application of this measurement to an expressive French database, shows that the articulation degree depends on expressivity. Using Lindblom's theoretical framework, we have tried to relate this articulation degree to the activation degree of a dimensional representation of emotions. Finally, speech processing is used to modify the articulation degree of a spoken or synthesized neutral utterance in order to apply a given expressivity to it.

1. Introduction

In our previous works, we constructed a database management system for speech, allowing the manipulation of large corpora for various artistic objectives. The first objective, was the statistical analysis of acoustic variables according to the context (phonetic, prosodic, stressing). This type of analysis allowed us to observe the influence of the expressivity on the speech rate [1]. The second objective lied in high quality Text-To-Speech synthesis. One of the purposes of this was the reconstruction of the voice of a specific speaker, for example that of a dead celebrity. An extension of this allowed musical synthesis and cross synthesis of speech and music using prosody [2] for composers of contemporary music. Currently, the system is being used to context-dependently transform the expressivity in speech, for film and theater directors and for dubbing studios. It is within this framework that the study is conducted.

We recorded four French actors performing a set of expressive sentences. *Expressivity* defines all the acted and simulated emotions, as well as the attitudes and moods that were part of the actors approach. This data was phonetically segmented and then analyzed focusing on several aspects : the pitch, intensity, speech rate, vocal quality and articulation degree. The study of these acoustic descriptors in relation to the five dimensions of the prosody [3] allowed us to observe tendencies for each of the expressivities. These tendencies were then used to transform the expressivity of a neutral sentence with speech processing algorithms. This study concerns specifically the analysis and modification of the articulation degree of various different expressivities.

Certain theories on emotions [4] commonly represent themselves in three dimensions. These dimensions are *valence* (positive - negative), *power* (weak or strong) and *activation*. The activation degree of an emotion conveys if the speaker is brought to act or to remain passive when they are experiencing a state of emotion. Subjects experiencing stress or under depression do not articulate voiced sounds with the same effort as subjects in a neutral emotional state [5]. The measure of the activation degree in emotional speech can be connected to the articulation degree theory of Lindblom [6]. The "H and H" theory proposes two articulation degrees of speech : *Hyper* speech which exhibits maximal clarity and *Hypo* speech which is produced in the most economic way possible.

This article proposes a novel measure of Lindblom's articulation degree. It comes from an attempt to provide an acoustic measurement of the activation level in the case of expressive speech. First, it details analysis methods developed to estimate the formant trajectories and the articulation degree. Then, this measurement is applied to an expressive French corpus. Finally, signal processing techniques are used to modify the articulation degree of a neutral utterance and to apply a desired expressivity to it.

2. Articulation Degree Measurement

The articulation degree originates from interactions between the phonetic context, the speech rate, and the spectral dynamics (which corresponds to the speed of the change in the configuration of the vocal tract). So the traditional measurement [7] of the articulation degree consists of defining formant targets for every phoneme, by taking into account coarticulation phenomenon, and studying the differences between the realizations and the targets with regard to the speech rate. Considering the difficulty of defining the targets, we opted for a statistical measurement of the articulation degree.

2.1. Prerequisite

The proposed measurement of the articulation degree of expressivity results from the joint observation of the evolutions in the area of the vocalic triangle and the speech rate. Thus it requires beforehand three types of speech signal analysis. First, sentences must be phonetically segmented to know which phonetic category belongs to every portion (frame) of the signal. Then, the syllabic structure built over the previous phonetic segmentation, allows the dynamic measurement of the local speech rate [1]. Finally, with the estimation of the formant trajectories, the measurements of the vocalic triangle can be defined.

2.2. Formant frequency estimation

There exists numerous tools that estimate the frequency of formants [8]. The majority of these tools model the spectral envelope of the segmented and windowed signal, by an autoregressive system, the poles P of which correspond to the resonances of the vocal tract. By filtering these poles according to

importance criterias, they define for every frame n/N, a restricted set of candidate poles $P_{can}(n)$ among which some correspond to formants. But these tools do not attribute a formant index to the poles. They look at a set of possible candidates but they do not assign each candidate to a particular formant. However, the attribution stage of the poles to the formants is necessary for the observation and modification of the vocalic triangle because this requires the knowledge of the 1^{st} (F_1) and of the 2^{nd} (F_2) formant frequencies.

Traditionally, the affectation of the sets $P_{can}(n)$ to the formants can be changed by the user. This assembles the poles according to their frequency ranges. For example, a male voice, with a pole where the frequency is situated between 800Hz and 1500Hz is often indexed as the 2^{nd} formant. But this a priori can skew the results if several formants are present in a frequential area or if the frequency of a formant exceeds these limits, which is sometimes the case in the expressive speech (see figure 1. Moreover, the quantity of data used requires an automation of the task.

2.3. Formant-Viterbi algorithm

We present a new algorithm called formant-Viterbi which allots to each formant, one pole's candidates, without an a priori on its frequential area, and which simultaneously takes into account a constraint of continuity on the formant trajectories.

2.3.1. Hypothesis

This algorithm is based on three assumptions :

- Hyp_1 : formants correspond to the prominent poles of the spectral envelope modeled with an autoregressive system.
- Hyp_2 : these poles can be classified according to their prominence and their respective positions in comparison to each other.
- Hyp_3 : the trajectory of a formant has a certain continuity in the time-frequency space.

The constraint of continuity in the trajectories allows a decrease in noise between estimations from one frame to the other.

2.3.2. Assignment of a pole to a formant

The first stage is a quasi-derivation of the N-framed and windowed signal. Then a predictive linear analysis (LP) of the filtered signal is carried out. For each frame n, the roots of the polynomial constitute the P poles of the spectral envelope. For each pole p of a frame n, one measures :

- $F(p, n)$: the frequency of the pole (phase)
- $Q(p, n)$: the bandwidth (proximity of the pole to the unit circle)
- $Gd(p, n)$: the group delay of the LP polynomial at the frequency of the pole [9]
- $A(p, n)$: the amplitude of the LP polynomial at the frequency of the pole

The three last parameter values of the poles are normalized in relation to the temporal horizon corresponding to the sentence. A weight (between 0 and 1) is allotted to each one of these parameter values. For the frame n, the probability of membership of a pole p to a formant $P(p, n)$, is given by the weighted sum of these parameter values (Hyp_1). The matrix $P(p, n)$ represents a whole sentence in the time-frequency space. At a given moment n, $P(p, n)$ calculates the probability that a formant trajectory will take a pole frequency f_p.

2.3.3. Formant trajectory

The constraint of the spectro-temporal continuity of formant trajectory (Hyp_3) is represented by the transition probability matrix $T(f_p, f_p)$ of Toeplitz (symmetrical and circular). The trajectories of the formants are "decoded" recursively, one after the other, by a Viterbi algorithm which takes into account the N frames of the sentence. Dynamic programming makes it possible to draw a formant trajectory on the matrix $P(p, n)$ while respecting continuity constraint $T(f_p, f_p)$ between each frame. The trajectory of the first formant is estimated by initializing its frequency with 0Hz at the first frame ($t = 0$). After a first decoding iteration, poles where the frequency is below the first formant trajectory are eliminated from $P(p, n)$ (Hyp_2). Then the trajectory of the second formant is decoded in the same way and so on.

2.4. Formant frequency of a phone

These estimated trajectories make it possible to know the formant frequency evolutions throughout each phone. In order to minimize the estimation errors and to obtain only one representative value per phone called *characteristic frequency*, a local statistical measure is carried out on all the temporal frames of a phone, by using the phonetic segmentation. A 2-order polynomial of Legendre, models the temporal evolution of the trajectory of a formant through the phone. If the evolution of the frequency is linear, the *characteristic frequency* corresponds to the median value of the frequencies taken on frames bordering the middle of the phone. If the evolution of the frequency is parabolic, which shows that the formant reached a target then deviated from it (coarticulation), the *characteristic frequency* corresponds to the median value of the frequencies taken on the frames bordering the moment when the frequency reaches its target (moment when the derivative of the 2^{nd} order polynomial nullifies). This measure reflects better the intended meaning by the speaker during the pronunciation of the vowel and possesses a lower variance than that of the average calculated on all frames of the phone.

2.5. Vocalic triangle

Vocalic triangle is the name given to the geometrical figure which form the vowels /a/, /i/ and /u/ (see figure 1), when they are placed in a bi-dimensional space called *cardinal space*, the axes of which are F_2 and F_1.

Each of these bordering vowels, with a given expressivity and a given expressive power, are assigned the average statistics of the *characteristic frequencies* of all phones corresponding to their phonetic class. For ten sentences measured by expressivity and power degree, each of these averages imply about thirty individuals. That is why we also represent the variances of the measurement (widths of ellipsoids). It has to be noted that possible variabilities due to coarticulation are avoided since every vocalic triangle is estimated on vowels situated in the same phonetic context (see part 3.1).

3. Influence of the expressivity

A study concerning the influence of the segment duration on the vocalic triangle in neutral speech [10], shows that formants aim towards a central vowel for segments of short duration, resulting in a diminishing of the vocal triangle area. In the neutral case, an acceleration and a deceleration correspond respectively to a reduction and an expansion of the vocalic tri-

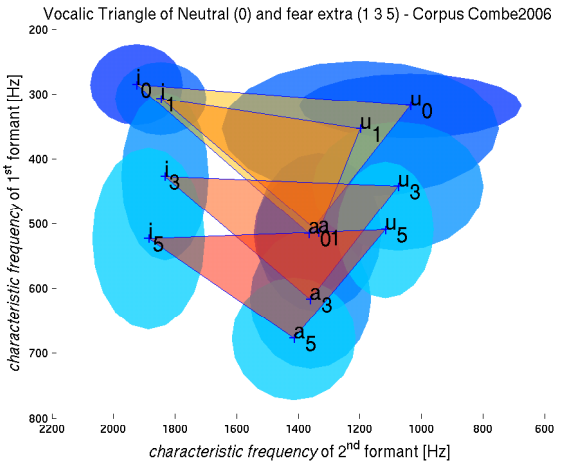

FIG. 1 – Vocalic triangles in the neutral case (a_0 ; u_0 ; i_0) and in the case of extrovert fear with three different power degrees (weak 1, average 3 and strong 5). Center and widths of ellipsoids represent respectively means and variances.

angle. This phenomenon is called NTN, Natural Tendency of the Neutral case for continuation. A major difference exists between neutral and other expressivities : Given a phonetic context, the articulation degree is not only dependent on the speech rate variable but also on expressivity.

3.1. Expressive corpus

The *IrcamCorpusExpressivity* database is composed of recordings of four actors, two males and two females, recorded during approximately one hour and a half each. They were all recorded in the same professional conditions following an identical procedure. Ten sentences extracted from a phonetically balanced corpus [11] have been marked with prosodic boundaries using punctuation and indicated stresses on specific syllables of words. Chosen *expressivities* were acted emotions : *neutral, introvert and extrovert anger, introvert and extrovert happiness, introvert and extrovert fear, introvert and extrovert sadness*, as well as *positive* and *negative surprises*, *disgust*, *discretion*, *excitation* and *confusion*. Each sentence was pronounced with all *expressivities*. Furthermore, in the case of acted emotions, every sentence was repeated six times with an increasing expressive power degree. Finally, the corpus is composed of approximately 550 phonetically labeled utterances per actor. Some *fillers* have been also uttered for each *expressivity*.

3.2. Power degree and vocalic triangle

The figure 2 presents four vocalic triangles superimposed and measured. In the neutral case (the vowels of which are indexed by 0) and in the case of extrovert fear with three different power degrees (1,3,5) the vowels are represented by ellipses among which the coordinates of the center and the widths are respectively defined by the means and the variances of the *characteristic frequencies* of the 2^{nd} and 1^{st} formants (see part 2.5). This figure shows a geometrical translation of the vocalic triangle as the power degree increases. It is partly explained by a higher pitch (rising according to expressive power degree) than the neutral case (see figure 2). What it does not show, is that the speech rate increases according to the expressive power degree. Regarding the NTN, a significant reduction of the area is expected, since the speech accelerates, and is not observed (non significantly according to variances).

3.3. Expressivity and articulation degree

The NTN is clearly not designed for expressivities like extrovert sadness, introvert fear, and surprises which show a reduction of the vocalic triangle despite slower speech. Conversely, extrovert anger shows an expansion of the vocalic triangle area which exceeds that of the NTN. The use of the Nyquist criterion for distribution separation on the proposed [vocalic triangle area/speech rate] space clearly separates expressivities. For a given expressivity, the measurement of the articulation degree is explicitly connected to its position in comparison to the neutral (reference).

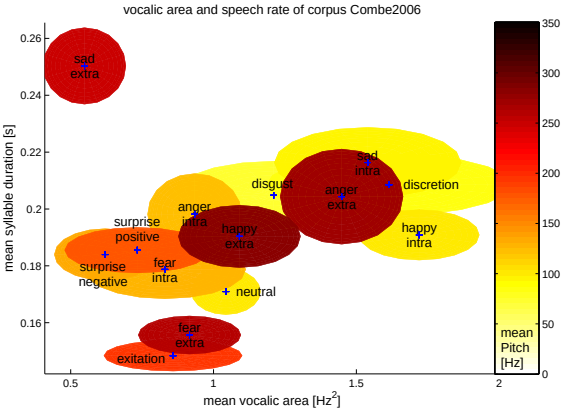

FIG. 2 – Expressivities performed by a male actor, represented according to their vocalic area (X axis), to their mean syllabic duration (Y axis) and to their mean pitch (Z axis - colors). Ellipsoids show mean values (center coordinates) and variances (widths)

3.4. Activation degree

The activation degree of an emotion specifies if the speaker reacts or remains passive when they are in an emotional state. A prior supposition is that the expressivities of the negative activation degree (passivity) and the positive (activity) are reliable with respect to the introversion and extraversion of the speaker. Thus, the activation degree could be related to the hypo- and hyper-articulated conditions of Lindblom's theory. The resulting measures do not allow confirmation of this supposition as we observed different strategies for the same expressivity depending on the actor (especially depending on the sex of the actor). To confirm, it would be necessary to confront results with perceptive tests and to involve spontaneous emotional data. On the other hand, it seems that all the actors performed the extrovert emotions by speaking about one or two octaves higher.

4. Transformation of the articulation degree

Our aim is to transform a given *neutral* utterance (source) by applying a chosen expressivity *E* and by a specific power degree *D* (target). Using phonetic, stressing and expressivity levels, two corresponding acoustic descriptors sets are predic-

ted by a generative statistical model [12], one using neutral expressivity (source) and the other using desired *expressivity E* (target). Inferred formant *characteristic frequency* distributions are then compared so as to provide a dynamic Frequency Warping Function $FWF(f,t)$. The joint acoustic modification of the prosodic dimensions results in the transformation of the articulation degree.

Dynamic frequency warping algorithms possess numerous applications in speech processings [13]. Speaker normalization/adaptation for speech recognition and voice conversion needs VTLN (Vocal Tract Length Normalization) which can be achieved by frequency warping. Here it is employed in order to move the formant frequencies as described in [14]. At moment t_0, the frequency axes of the source spectral envelope and of the target spectral envelope are linked by the piece-wise linear function $f_{target} = FWF(f_{source}, t_0)$. It is designed by the linear interpolation of the break point function generated by the model using formant frequencies of the source and of the target. Sampling of the source spectral envelope (amplitude S_s and unwrapped phase ϕ_s) by the resulting non-linear target frequency axis provides a target envelope (S_t and ϕ_t).

$$\begin{cases} S_t(Fs,t) & = S_s(FWF(Fs,t),t) \\ \phi_t(Fs,t) & = \phi_s(FWF(Fs,t),t) \end{cases} \quad (1)$$

Source spectral envelope is transformed into target spectral envelope thanks to a phase vocoder technology [15]. Some examples can be heard at the following address : `http://www.ircam.fr/anasyn/beller`.

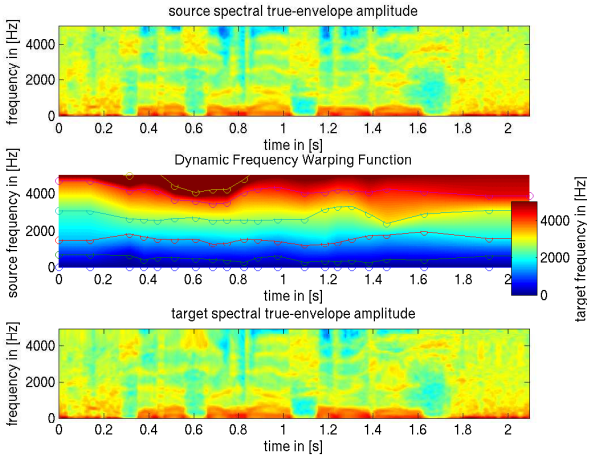

FIG. 3 – Example of source spectral envelope modification using dynamic frequency warping

5. Conclusion

In this paper we presented our motivations for the analysis/modification of the articulation degree in expressive speech. After the presentation of an algorithm for the estimation of the formant trajectories, we described a measurement of the articulation degree involving a statistical study of the vocalic triangle area and of the speech rate, according to the expressive power degree. The application of this measurement to an expressive French corpus highlighted the influence of the expressivity on the articulation degree. Possible links between activation degree and articulation degree have been discussed. Finally, a speech processing algorithm was used to transform the degree of articulation in a neutral sentence so as to confer a desired expressivity to it.

6. Acknowledgments

This work was partially funded by the French RIAM network project VIVOS.

7. References

[1] Grégory Beller, Diemo Schwarz, Thomas Hueber, and Xavier Rodet, "Speech rates in french expressive speech," in *Speech Prosody*, Dresden, may 2006, SproSig, ISCA.

[2] Grégory Beller, Diemo Schwarz, Thomas Hueber, and Xavier Rodet, "Hybrid concatenative synthesis in the intersection of speech and music," in *JIM*, Anne Sedes et Horacio Vaggione, Ed., 2005, vol. 12, pp. 41–45.

[3] H.R. Pfitzinger, "Five dimensions of prosody : Intensity, intonation, timing, voice quality, and degree of reduction," in *Speech Prosody*, H Hoffmann, R. ; Mixdorff, Ed., Dresden, 2006, number 40 in Abstract Book, pp. 6–9.

[4] Marc Schröder, "Expressing degree of activation in synthetic speech," in *IEEE Transactions on Audio, Speech and Language Processing*, July 2006, vol. 14, pp. 1128–1136.

[5] F.J. Tolkmitt and K.R. Scherer, "Effect of experimentally induced stress on vocal parameters," *J. Exp. Psychol. [Hum. Percept.]*, vol. 12, no. 3, pp. 302–313, 1986.

[6] Bjorn lindblom, *Economy of Speech Gestures*, vol. The Production of Speech, Spinger-Verlag, New-York, 1983.

[7] J. Wouters and M. Macon, "Control of spectral dynamics in concatenative speech synthesis," in *IEEE Transactions on Speech and Audio Processing*, 2001, vol. 9, pp. 30–38.

[8] P. Boersma, "Praat, a system for doing phonetics by computer," in *Glot international*, 2001, vol. 5 of *10*, pp. 341–345.

[9] H.A. Murthy, K.V. Madhu Murthy, and B. Yegnanarayana, "Formant extraction from phase using weighted group delay function," in *Electronics Letters*. 1989, vol. 25, pp. 1609–1611, IEE.

[10] Cédric Gendrot and Martine Adda-Decker, "Analyses formantiques automatiques de voyelles orales : évidence de la réduction vocalique en langues française et allemande," in *MIDL*, 2004.

[11] Pierre Combescure, "20 listes de dix phrases phonétiquement équilibrées," *Revue d'Acoustique*, vol. 56, pp. 34–38, 1981.

[12] Grégory Beller, "Context dependent transformation of expressivity in speech using a bayesian network," in *ParaLing*, Saarbr ucken, August 2007.

[13] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Frequency-warping in speech," in *ICSLP*, Philadelphia, PA, 1996, vol. 1, pp. 414–417.

[14] Zhi-Wei Shuang, Raimo Bakis, Slava Shechtman, Dan Chazan, and Yong Qin, "Frequency warping based on mapping formant parameters," in *INTERSPEECH*, 2006, number 1768.

[15] Niels Bogaards, Axel Roebel, and Xavier Rodet, "Sound analysis and processing with audiosculpt 2," in *International Computer Music Conference (ICMC)*, Miami, USA, Novembre 2004.