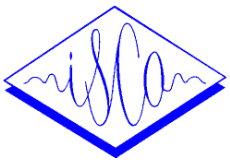

VOQUAL'03, Geneva, August 27-29, 2003

Emotions and Voice Quality: Experiments with Sinusoidal Modeling

*Carlo Drioli, Graziano Tisato, Piero Cosi, and Fabio Tesser**

Laboratory of Phonetics and Dialectology
ISTC-CNR, Institute of Cognitive Sciences and Technology, ITALY
{drioli,tisato,cosi}@csrf.pd.cnr.it

*Centre for Scientific and Technological Research
ITC-IRST, ITALY
tesser@itc.it

Abstract

Voice quality is recognized to play an important role for the rendering of emotions in verbal communication. In this paper we explore the effectiveness of a sinusoidal modeling processing framework for voice transformations finalized to the analysis and synthesis of emotive speech. A set of acoustic cues is selected to compare the voice quality characteristics of the speech signals on a voice corpus in which different emotions are reproduced. The sinusoidal signal processing tool is used to convert a neutral utterance into emotive utterances. Two different procedures are applied and compared: in the first one, only the alignment of phoneme duration and of pitch contour is performed; the second procedure refines the transformations by using a spectral conversion function. This refinement improves the reproduction of the different voice qualities of the target emotive utterances. The acoustic cues extracted from the transformed utterances are compared to the emotive original utterances, and the properties and quality of the transformation method are discussed.

1. Introduction

The transmission of emotions in speech communication is a topic that has recently received considerable attention. Automatic speech recognition (ASR) and text-to-speech (TTS) synthesis are examples of popular fields in which the processing of emotions can have a substantial impact and can improve the effectiveness and naturalness of the man-machine interaction. Many of the researches in the field have emphasized the importance of prosodic features (e.g., speech rate, intensity contour, F0, F0 range) and the importance of the voice quality in the rendering of different emotions in verbal communication [1, 2, 3]. In TTS technologies, voice processing algorithms for emotional speech synthesis have been mainly focusing on the control of phoneme duration and pitch, which are the principal parameters conveying the prosodic information. On the side of voice quality transformations for speech synthesis, some recent studies have addressed the exploitation of source models within the framework of articulatory synthesis to control the characteristics of voice phonation [4, 1].

The aim of this paper is to explore the effectiveness of a sinusoidal modeling processing framework for voice transformations finalized to the analysis and synthesis of emotive speech. In the field of TTS, the sinusoidal modeling approach is appreciated for providing a flexible signal representation that can be used to implement many signal processing tasks, such pitch- and time-scale modification, with good quality results [5]. Recently, even more sophisticated transformations have been proposed, such as transformation of spectral features for speaker conversion [6, 7, 8]. This last approach is used in the present paper to evaluate the effectiveness of the sinusoidal based voice conversion approach when used to reproduce the voice quality differences that characterize different emotions.

The paper is organized as follows. In Section 2 the voice material is introduced and the principal acoustic cues considered are described. Results on the discriminant ability of cues are also reported. In Section 3 we define a signal processing framework, based on a sinusoidal representation, for transforming the prosodic and voice quality characteristics of speech. The framework is evaluated on a set of examples from the database and the characteristics and limitations of the method are discussed.

2. Voice Material

A male University student, who speaks a northern regional Italian and with recitation skills, pronounced two phonological structures 'VCV, corresponding to two feminine proper names: "Aba" /'aba/ and "Ava" /'ava/, simulating, on the basis of appropriate scenarios, six emotional states: anger (A), joy (J), fear (F), sadness (SA), disgust (D) and surprise (SU), apart from the neutral one (N), corresponding to a declarative sentence. This 14 words set was repeated many times in random order. For each series of recordings the rest position has been recorded as well. In the same session we collect the articulatory data with an optotracking 3D movement analyzer system (ELITE), which allows a synchronous recording of the acoustic signal. This system ensures high accuracy (100 Hz sampling rate, maximal error of 0.1 mm for a 28x28x28 cm cube) and minimum discomfort to the subject because it tracks the infrared light reflected by small (2 mm diameter), passive markers glued on external lips contour and on the face. The data collected have been used to analyse the complex interactions between the articulatory movements, due to the phonetic-phonological constraints, and the face configurations, due to the emotions [9].

In the following we report a brief description of the voice material in terms of acoustic cues commonly related to emotions. Table 1 shows the mean values for duration, F0, F0 range, and intensity of the stressed vowels in both words /'aba/ and /'ava/. An in-depth statistical analysis (ANOVA) of these and other acoustic and articulatory parameters, has been performed for the same voice corpus in a companion study [9].

Anger (A) was characterized by the highest intensity, mid-range F0, narrow F0 range, and shorter duration than the neutral (N). If we express in musical terms the anger to neutral F0 ratio, this is precisely a 7/5 interval (tritone). Both the stressed and unstressed vowels were characterized by a harsh quality, with a clear predominance in the stressed vowel. Disgust (D) had longest duration, mid-range intensity, mid-range F0, narrow F0 range, and the disgust to neutral F0 ratio is approximately a major second. Stressed and unstressed vowels of D were characterized by a creaky voice quality, slightly more pronounced in the unstressed vowel. Joy (J) and surprise (SU) both presented high F0 and wide F0 range only in the stressed vowel, medium-high intensity, and shorter durations than the neutral in both stressed and unstressed vowels. The emotive to neutral F0 musical ratio for these two emotions approximate an augmented octave. These two emotions where the most difficult to distinguish perceptually. Both presented a "bright"-sounding quality. A distinctive quality aspect of joy was a breathy voice onset in the stressed vowel, whereas the surprise had a sharp onset. Fear (F) presented mid-range duration, highest pitch and low pitch range, mid-range intensity. The fear to neutral F0 ratio is a major second interval. A breathy voice onset in the stressed vowel, and a general breathy voice quality, characterized this emotion. Sadness (SA) had high F0 and wide F0 range values, mid-range intensity and duration. No distinctive voice quality features could be observed, other than an overall "dark"-sounding quality. A general interesting observation, suggested by Table 1, is that anger, disgust, fear and sadness present an evident dissonant interval for what concerns the emotive to neutral F0 mean ratio.

Table 1: Mean values for duration, F0, F0 range, and intensity of the stressed vowels in /'aba/ and /'ava/.

	Duration (s)	F0 (Hz)	F0range (Hz)	Intensity (dB)
A	0.195	177.744	18.276	76.735
D	0.293	138.993	14.935	72.297
N	0.231	126.428	14.588	70.819
J	0.188	260.999	67.873	74.928
F	0.211	288.737	26.889	70.829
SU	0.179	265.867	89.857	72.527
SA	0.269	209.032	56.582	70.415

2.1. Voice quality indexes and statistical analysis

The speech signal has been manually segmented and analysed by means of a voice analysis software (PRAAT [10]) and of Matlab routines. The following set of cues, which are among the ones that are most commonly found in investigations on emotive speech, have been selected as voice quality correlates of emotions [11, 12]: *Shimmer* and *Jitter*, i.e. the cycle-to-cycle variations of waveform amplitude and fundamental period respectively; the Harmonic-to-Noise ratio (*HNR*), defined as the ratio of the energy of the harmonic part to the energy of the remaining part of the signal; the Hammarberg Index (*HammI*), defined as the difference between the energy maximum in the 0-2000 Hz frequency band and in the 2000-5000 Hz band; the drop-off of spectral energy above 1000 Hz (*Do1000*), computed as the gradient of the least squares approximation of the spectral slope above 1000 Hz; the relative amount of energy in the high- (above 1000 Hz) versus the low-frequency range (up to 1000 Hz) of the voiced spectrum (*Pe1000*); a spectral flatness measure (*SFM*), computed as the ratio of the geometric to the arithmetic mean of the spectral energy distribution.

The acoustic cues were computed in the stressed and unstressed vowel segments for each recording. Moreover, a group average of the cues for each emotion was computed. The standardized difference to the neutral for each acoustic cue is reported for the stressed vowel and the unstressed vowel in /'aba/ (Fig. 1) and /'ava/ (Fig. 2). The cue values are grouped in the plots so to define an acoustic profile for each emotion. The cue profiles for the stressed and unstressed vowel are characterized by significant differences in the values for the different emotions.

A few comments are worth noticing for some remarkable cases: the high *Shimmer* value for anger (A) reflects the harsh nature of the voice during a relevant portion of the vowel /'a/. From an acoustic point of view, a short-time frequency analysis reveals the presence of a subharmonic component in the mid range of the spectrum, which produces a fast amplitude modulation in the time-domain. Disgust (D) presents an almost flat pattern in the stressed vowel, whereas it noticeably differs in *Shimmer*, *Jitter*, and *SFM* in the unstressed vowel, in agreement with its emphasized creaky voice quality. Jot (J) and surprise (SU) present the patterns with more similarities, especially in the unstressed vowel where they differ appreciably only for the *SFM* parameter. Actually joy and surprise were the most difficult to distinguish one from the other also by informal perceptual listening. In fear (F) the low values of *Pe1000* is due to the extremely high pitch, which is also kept stable during the vowel, as opposed to what happens for joy (J) and sadness (S), in which the pitch ranges from very low to very high values, affecting the average.

A linear discriminant analysis was performed on the voice data set. The percentage of correct classification by the Jackknifed procedure for the stressed and unstressed vowel attained respectively 60% and 65% (the classification matrices are reported in Tables 2 and 3).

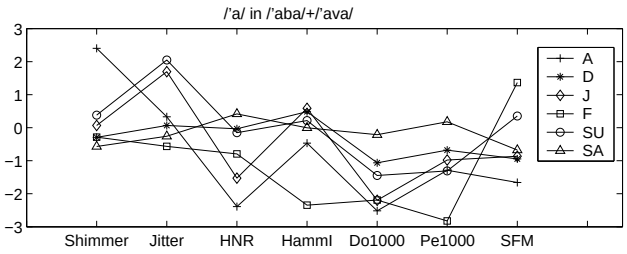

Figure 1: Patterns of audio cues for the different emotions (stressed vowel segment in the words / 'aba/ and / 'ava/).

Table 2: Jackknifed classification matrix for the stressed vowels

	A	D	N	J	F	SU	SA	% correct
A	11	0	0	3	0	0	0	79
D	0	7	3	1	0	0	3	50
N	0	4	12	0	0	0	6	55
J	0	3	0	14	2	3	0	64
F	0	1	0	0	13	0	0	93
SU	0	4	0	5	4	9	0	41
SA	0	2	5	0	0	0	7	50
Total	11	21	23	19	12	16	18	60

From the classification matrix it is evident that the best recognition performance is attained for fear, anger and joy. The

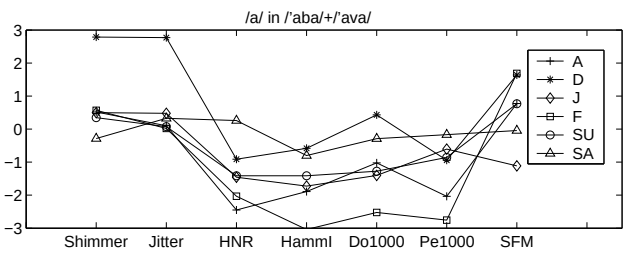

Figure 2: Patterns of audio cues for the different emotions (unstressed vowel segments in / 'aba/ and / 'ava/).

Table 3: Jackknifed classification matrix for the unstressed vowel

	A	D	N	J	F	SU	SA	% correct
A	10	0	0	1	1	1	0	77
D	2	8	3	0	0	0	1	57
N	0	0	14	1	0	1	6	64
J	0	1	0	16	1	4	0	73
F	0	0	0	0	11	3	0	79
SU	1	0	2	4	3	11	1	50
SA	0	0	2	2	0	1	9	64
Total	13	9	21	24	16	21	17	65

worst performance is attained for surprise and disgust. From the observation of the canonical discriminant functions for the stressed vowel (not reported here) we were able to say that the first factor was responsible alone for the discrimination of anger (A) from the other emotions, and had the *Shimmer* and *Jitter* cues as important components. This is in agreement with Fig. 1 and with the nature of the harsh quality of the stressed vowel. On the other hand, for the unstressed vowel, none of the factors were able to discriminate a particular emotion from the other. Since from Fig. 2 disgust is the only emotion characterized by high differences in *Shimmer* and *Jitter*, these cues have not strong relevance in the configuration of the canonical factors, and the recognition score for disgust is among the worsts.

3. Neutral to emotive utterance mapping

The investigation relies on the well known sinusoidal model of the signal [13]. The analysis algorithm acts on windowed portions (*frames*) of the signal, and produces a time-varying representation as sum of sinusoids (here called *partials*. Assuming that the number of partials H is constant for all frames, for the ith frame the result of the sinusoidal modeling is a set $\{(f_h(i), a_h(i), \phi_h(i)), h = 1, \ldots, H\}$ of triples of frequency, magnitude and phase parameters describing each partial, and a residual noise component. H is taken sufficiently high to provide the maximum needed bandwidth, and zero magnitude is assigned to the exceeding partials for the spectra with lower bandwidth. The re-synthesis of sound relies on the inversion of the analysis procedure, i.e. on inverse transformation of the sinusoidal analysis and on overlap-and-add of time-domain frames. The residual noise is not considered in the present study, due to the lack of a reliable noise model. Our experience is that the inclusion of an inadequately modeled noise introduces undesirable artifacts and significantly degrades the result of the synthesis. Various approaches for the representation of noise have been used in speech processing [14, 5], and the inclusion of this component will be the subject of future investigations.

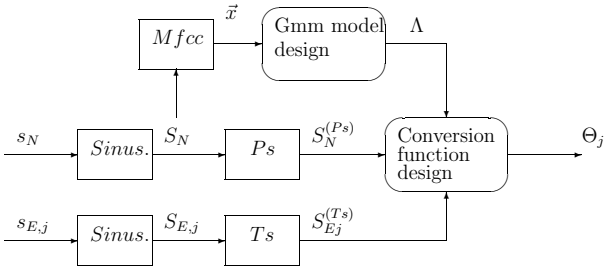

Figure 3: Scheme of the conversion function design. The parameters $\mathbf{\Theta}_j$ represent the spectral map for the conversion neutral $\rightarrow E_j$ (jth emotion).

The sinusoidal representation allows to control some of the basic speech signal characteristics, such as timing, pitch, and intensity, by simply interpolating analysis frames, and by shifting or scaling the frequency and magnitude of the partials. When performing pitch variations for speech signals, it is common practice to interpolate the shifted frequencies with respect to the original spectral envelope so that the formants are maintained. Unfortunately, no such simple rules are available in general for reproducing the voice quality transformations implied in the production of different emotions in speech. We thus rely here on a statistical approach which permits to learn the spectral transformations from a database of emotive utterances. The spectral processing method uses a GMM-based mapping function trained on spectral data from the voice corpus [6]. The conversion function used has the form

$$\mathcal{F}(\vec{x}_t) = \sum_{i=1}^{M} p(\lambda_i|\vec{x}_t)[\vec{\theta}_i], \tag{1}$$

where $\vec{x}_t$ is a spectral representation of frame t (we use here the mel-frequency cepstral coefficients, Mfcc), and $\mathbf{\Theta} = [\vec{\theta}_1 ... \vec{\theta}_M]$, is the set of parameters of the mapping function. The term $p(\lambda_i|\vec{x}_t)$ is the probability that an input acoustic vector $\vec{x}_t$ belongs to the class $\lambda_i = (\alpha_i, \vec{\mu_i}, \mathbf{\Sigma_i})$. The gaussian mixture is completely specified by the mean vectors, covariance matrix and mixture weights, and can be represented by

$$\Lambda = \{\alpha_i, \vec{\mu_i}, \mathbf{\Sigma_i}\}\ i = 1 \ldots M \tag{2}$$

An acoustic model Λ is computed for the neutral utterance. Given the sequence of T training vectors $\mathbf{X} = \{\vec{x}_1, ...\vec{x}_T\}$, this is achieved by the the ML estimate computed using the expectation-maximization (EM) algorithm, which maximizes the GMM likelihood $p(\mathbf{X}|\Lambda) = \prod_{t=1}^{T} p(\vec{x}_t|\Lambda)$. The design of spectral mapping functions requires a preliminary time- and pitch-alignment between the neutral utterance and the other emotive utterances. Let say that J is the number of emotions,not considering the neutral. After the time- and pitch- alignment step is completed, a mapping function $\mathbf{\Theta}_j$, $j = 1, ..., J$ is computed for each emotion. The conversion function is designed so as to add to the pitch-shifted spectral envelope from the neutral at frame t and to reproduce the desired spectral envelope from the time-aligned emotive utterance:

$$\tilde{S}_{Ej,t}^{Ts} = \mathcal{F}_j(\vec{x}_t) + S_{N,t}^{Ps}, t = 1, ..., T, \tag{3}$$

where $\tilde{S}_{Ej,t}^{Ts}$ is the reproduced target spectrum envelope after time-alignment, and $S_{N,t}^{Ps}$ is the spectrum envelope obtained from the neutral by pitch shifting. Note that a sequence of

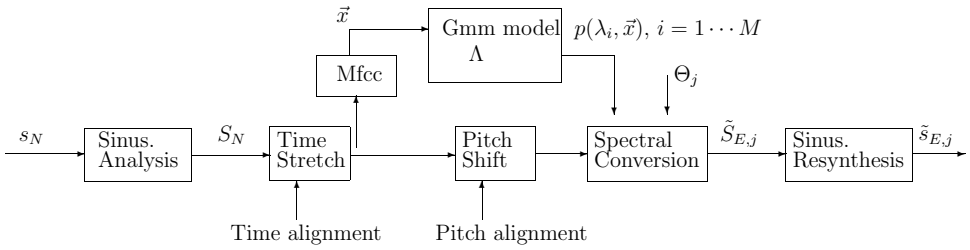

Figure 4: Scheme of speech transformations to adapt neutral utterance to emotive utterance

frames in a region with approximately same pitch could happen to be mapped into a sequence of target frames from a region where the pitch has wide variations. In order to take this into account, a term $\Delta p, t$ (the pitch shift factor at frame t) can be included in the design of the conversion function. The training procedure is schematized for the jth emotion in Fig. 3.

In order to highlight the differences between the processing in which no spectral conversion is performed and the refined processing with spectral conversion, two transformation schemes were used:

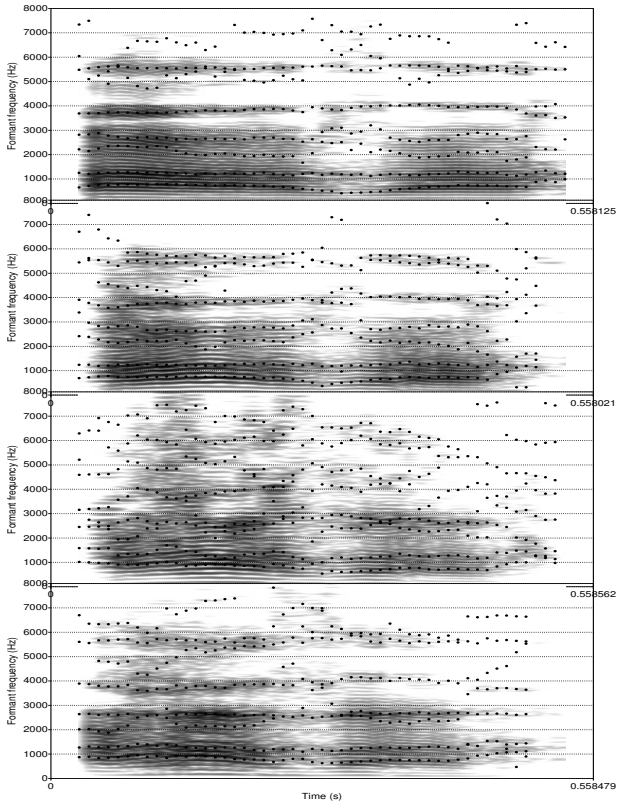

Figure 5: Result of the transformation neutral → anger. Upper panel: original neutral utterance. Second panel from top: neutral → anger (Ts+Ps). Third panel: neutral → anger (Ts+Ps+Sm). Lower panel: original anger utterance

1) Time stretching (Ts) and (formant preserving) pitch shifting (Ps). This transformation process is used to align the timing and pitch contour of the neutral recording to the timing and pitch contour of other emotions. After time and pitch transformation of the neutral utterance, a new utterance is obtained which has the prosodic characteristics of the target emotion, and the voice quality derived from the neutral utterance (the only spectral processing being the formant preservation).

2) Spectral modifications (Sm) based on (1) is performed to align the voice quality characteristics of the modified utterance to those of the target emotion.

In Fig. 5 the spectrograms of the signals resulting from the neutral → anger transformation are shown.

4. Experimental results

The transformations described in the previous section were performed on one repetition of the recordings of the word /'ava/. The neutral recording was first transformed on a frame-by-frame basis with the time-stretch and pitch-shift formant preserving procedures, so to match the timing and the pitch profiles of each of the other six emotive recordings. The resulting speech signals were segmented and analysed in order to obtain for the transformed signals the same set of acoustic cues introduced before. A further set of six transformations was obtained by computing the spectral transformations based on (1), and the corresponding set of cues was computed. The standardized difference to the neutral for each acoustic cue is reported in Fig. 6 (only results concerning the stressed vowel are shown). Comparing the cues from the Ts+Ps transformations (line: _ _+) and those from the Ts+Ps+Sm transformations (line: __∗) with the cues from the target samples (line: __o), we can see that in general Ts+Ps processing fails to reproduce the different acoustic patterns of the emotions. In most cases, the new values of the cues lie around the zero, meaning that there is little variation with respect to the values from the neutral utterance, except for an increase in the *HNR* values, since the residual noise is not modeled in the resynthesis, and a decrease in some of the cues related to spectral slope, probably due to the rising of the pitch. Perceptually, the utterances obtained by this transformation (Ts+Ps) are quite convincing since the sinusoidal processing framework preserves the naturalness of the original signal, and the comparison in terms of voice quality with the original target utterances is more effective since only the voice timbre differences are left.

The inclusion of the spectral processing stage in the procedure (Ts+Ps+Sm) leads to positive effects for a subset of cues, i.e., *HammI*, *Do1000*, *Pe1000*, *SFM*, whereas no benefits seem to be achieved for *Shimmer*, *Jitter*, and *HNR*. This, for *Shimmer* and *Jitter*, is motivated by the fact that the sinusoidal analysis/resynthesis framework, being based on a short-time fre-

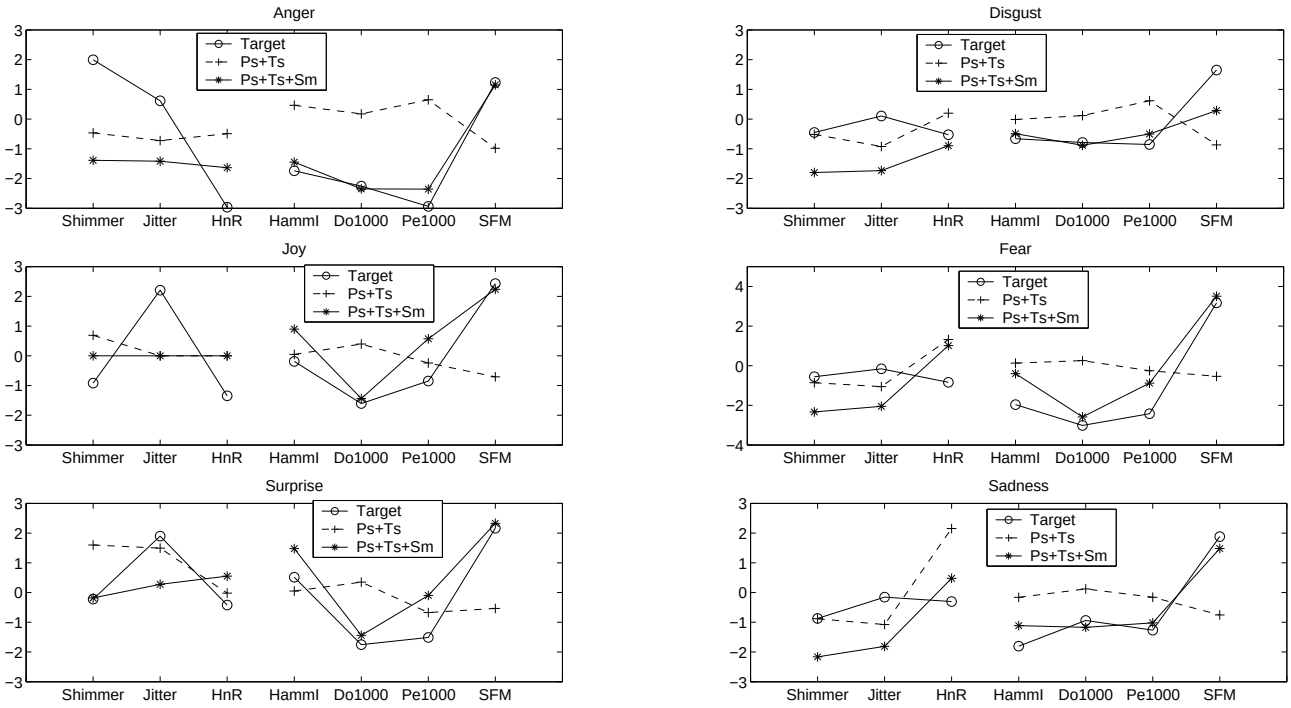

Figure 6: Results of speech transformations

quency representation and being not pitch-synchronous, fails to provide a good model for cycle-to-cycle amplitude and pitch variations; for *HNR*, this is due to the lack of a model for the noisy and other non-harmonic parts of the signal, that are not included in the resynthesis (it can be seen in the plots that the *HNR* calculated from the transformed signals is in general greater that the *HNR* calculated from the original targets). The benefits for the remaining cues can be easily explained considering that all of them represent spectral envelope features and are computed from the short-time spectrum of the signal. From the perceptual point of view, the utterances obtained by this transformation (Ts+Ps+Sm) are appreciably close to the target ones, the principal differences being due to the time resolution limits of the sinusoidal framework and to the absence of a noise model. In the reproduced anger (A), the high degree of roughness present in the original utterance is not perceived (and this is in agreement with the fact that *Shimmer*, *Jitter*, and *HNR* values are not well reproduced for anger, see Fig. 6. As noted before, roughness is characterized in the signal by a relevant noisy component and by an amplitude modulation which appears in the spectrum as a sub-harmonic series. The inclusion of such terms would probably improve the rendering of roughness in anger. In the reproduced joy (J) and fear (F), both characterized by a breathy voice quality, the absence of the noisy component is perceived in the attack portion of the stressed vowel. In the synthesized surprise (SU), the overall impression is that the transformation well reproduced the salient prosodic and voice quality characteristics. From informal subjective tests, SU resulted one of the better reproduced emotions.

5. Conclusions

The sinusoidal signal processing tool has been used to convert a neutral utterance into emotive utterances. Two different procedures have been applied and compared: in the first one, only the alignment of phoneme duration and of pitch contour is performed; the second procedure refines the transformations by using a spectral conversion function. This refinement improves the rendering of the different voice qualities peculiar to different emotions. The acoustic cues extracted from the transformed utterances have been compared to the emotive original utterances, and the results showed that some of the voice quality characteristics of the emotions could be reproduced with the proposed transformation method. Formal listening and recognition tests on the transformed utterances will be reported in a future work.

We can conclude that the sinusoidal framework for the analysis and resynthesis of emotive speech offers some desirable properties. What is missing is a set of ad-hoc modeling refinements which permit to accurately model some peculiarities of the voiced sounds found in emotive speech. The modeling of noise components, a well known topic in the field of speech synthesis, will be included in future investigations. The modeling of other peculiar characteristics, such as the cycle-to-cycle amplitude and pitch variations, or the presence of sub-harmonic series in the spectra, will be also addressed.

6. Acknowledgements

Part of this work has been sponsored by, PF-STAR (Preparing Future multiSensorial inTerAction Research, European Project IST- 2001-37599, http://pfstar.itc.it/).

7. References

[1] C. Gobl and A. N. Chasaide, "The role of the voice quality in communicating emotions, mood and attitude," *Speech Communication*, vol. 40, pp. 189–212, 2003.

[2] T. Johnstone and K. R. Scherer, "The effects of emotions

on voice quality," *in Proceedings of the XIV Int. Congress of Phonetic Sciences*, pp. 2029–2032, 1999.

[3] D. R. Ladd, K. E. A. Silverman, F. Tolkmitt, G. Bergmann, and K. R. Scherer, "Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect," *Journal of the Acoustical Society of America*, vol. 78, no. 2, pp. 435–444, August 1985.

[4] C. d'Alessandro and B. Doval, "Experiments in voice quality modification of natural speech signals: the spectral approach," *in Proceedings of the 3rd ESCA/COCOSDA Int. Workshop on Speech Synthesis*, pp. 277–282, 1998.

[5] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 1, pp. 21 –29, January 2001.

[6] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, March 1998.

[7] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 285–288, 1998.

[8] A. Kain and M. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 285–288, 2001.

[9] E. Magno Caldognetto, P. Cosi, C. Drioli, G. Tisato, and F. Cavicchio, "Coproduction of speech and emotion: bimodal audio-visual changes of consonant and vowel labial targets," in *submitted to AVSP2003*, S. Jorioz, France, September 2003.

[10] P. Boersma, "PRAAT, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[11] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *in Journal of Personality and Social Psycology*, vol. 70, no. 3, pp. 614–636, 1996.

[12] K. Alter, E. Rank, S. A. Kotz, U. Toepel, M. Besson, A. Schirmer, and A. D. Friederici, "Affective encoding in the speech signal and in event-related brain potentials," *Speech Communication*, vol. 40, pp. 61–70, 2003.

[13] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-34, no. 4, pp. 744–754, August 1986.

[14] M. W. Macon and M. A. Clements, "An enhanced ABS/OLA sinusoidal model for wavefor synthesis in TTS," *in Proceedings EUROSPEECH*, vol. 5, pp. 2327–2330, September 1999.