Natural transformation of type and nature of the voice for extending vocal repertoire in high-fidelity applications

Snorre Farner, Axel Röbel and Xavier Rodet

Ircam, 1 place Igor Stravinsky, FR-75004 Paris, France

Correspondence should be addressed to Snorre Farner (farner@ircam.fr)

ABSTRACT

Natural voice transformation will reduce the need for authentic voices in many situations, ranging from vocal services via education and entertainment to artistic applications. Transformation of one voice to correspond to that of another person has been studied for decades but still suffers from limitations that we propose to overcome by an alternative approach. It consists in modifying pitch, spectral envelope, durations etc. in a global way. While it sacrifices the possibility to attain a specific target voice, the approach allows the production of new voices of a high degree of naturalness with different sex and age, modified vocal quality (soft, breathy, and whisper), or another speech style (dullness and eagerness). The transformation of sex and age has been evaluated by a listening test.

1. INTRODUCTION

Speech is already widely used in video games and replaces gradually the use of written text. Speech material was initially recordings that were stored and replayed without modification when asked for in the game. This may be sufficient for a narrator's comments in a simple setting, but is now more and more replaced by text-to-speech synthesis, which offers the flexibility to make evolutive utterances by non-player characters (NPCs) at a lower cost. Also, speech synthesis allows the players to listen to what other players type, while the introduction of voice over Internet (VoIP) has made it possible for the players of multi-player games to communicate with each other by the simple use of their voice.

In addition, while video-game players for a long time have had the possibility to change the physical appearance of their characters, only recently are plug-ins available allowing the players to transform their voice in accordance with their character: MorphVOX, VA Voice Changer, and Fake Voice. These programs claim to be able to transform a man's voice to that of a woman and vice versa, to other human voices, or to a number of classic and innovative special-effect voices. They were not tested (Windows only), and the few sound examples provided were not enough for a test of the natural human transformations. No technical information about the transformation engine was found on their Internet sites although all are based on pitch and timbre transposition.

Obvious advantages of voice transformation in video games, whether for entertainment or educational purposes, have been mentioned, but high-quality natural voice transformation is also entering into artistic areas (e.g., theater and music) as well as into the film industry: transformation of the actors voice into another that is more suitable for the role, the use of one actor to achieve several voices in dubbing or animation movies, modification of accentuation of recorded speech, or creation animal voices, to mention a few applications. Combined with speech synthesis, the potential for applications is overwhelming in the other areas of entertainment, education, and business. Virtual animated speaking agents are appearing for supplying online customer service, and the new branch of automatic story telling is emerging.

Transformation of the voice of a given person (source voice) to that of another person (target voice), referred to as voice conversion, has been a dream put forth by numerous films, but despite persistent efforts in many research labs all over the world, this technology still suffers from reduced sound quality due to signal-processing artifacts and requires at present extensive parallel recordings of the source and target voices for the conversion to be feasible.

We propose an alternative approach, which, rather than trying to attain a specific target voice, favors the quality of the modified sound by controlling the transformation conditions. We show that it is nevertheless possible to change the identity of the source voice by changing its apparent size, sex and age, or making the voice breathy, softer, rougher or less happy, or even reducing it to a whisper.

The paper is organized into two parts: Secs. 2, 3, and 4 concern the technology to transform the voice while Sec. 5 reports the results of a perceptual evaluation of the transformation of sex and age.

2. DIFFERENCES BETWEEN VOICES

Apart from variation in natural *pitch range*, different voices are distinguished and recognized by their *timbre* (sometimes called "color") that depends on the *physiology* of the voice and the *phonatory settings*. The term timbre is often defined as the quality of a sound other than the pitch, duration, and loudness. For the voice we often use the term *voice quality* for grouping timbre-related qualities like dark, bright, soft, rich, noisy, pure, rough etc.

2.1. Voice physiology

Acoustically, the vocal organ consists of the vocal tract (mouth and nose cavities) as a resonance chamber and the larynx (the vocal folds (glottis), the false vocal folds and epiglottis) as the principal sound producing mechanism, thus called the *voice source*.

The specific configuration of the voice organ, such as the length and shape of the vocal tract and the vocal folds, varies from person to person and gives them their individual pitch range and timbre. Nevertheless, there are general differences depending on the sex and age of the person [1, 2, 3], although it might be difficult in some cases to guess the sex and age merely from the person's voice. The most important differences are the natural vibration frequency range of the vocal folds (perceived as *pitch* and measured as *F0*), the spectral distribution of the glottal source (for instance measured as *spectral tilt*), and the shape of the vocal tract (specific resonances and anti-resonances called *formants* and *anti-formants*).

Iseli et al. [3] have reported pitch means and ranges for male and female voices of ages ranging from 8 to 39 years: about 250 Hz for boys, decreasing to about 125 Hz from the age of 11 to 15 years, and about 270 Hz for girls, descending to about 230 Hz for adult women. Similar values were already published by Peterson and Barney [1] but without distinguishing boys and girls or specifying the age. However, they included average frequencies for the three first formants F1, F2, and F3 of men, women, and children for ten English vowels [1]. Averaging their formant frequencies over all vowels, we find that F1 increases about 14 % from a child voice to a woman's voice, and about 33 % to a man's voice. The increase is maybe slightly higher for F2, and about 18 % and 38 % for F3.

Finally, the aged voice presents a new set of characteristics: decreased intensity, breathiness, relatively high pitch (especially for men), lower flexibility, and perhaps trembling [4].

2.2. Phonatory settings

The voice can take many different phonatory settings (in addition to those necessary for making the phones of a language). For example, the vocal tract may be shaped to make a dark or bright color or sound nasal. More interesting for the current paper, however, are the possibilities of the larynx, which has a great repertoire of voice qualities.

Based on numerous studies by many researchers, John Laver has made a comprehensive discussion and summary of the phonatory settings of the larynx and their relation to the perceived voice quality [5]. He argues for the existence of six basic phonatory settings:

1. modal voice and falsetto (orthogonal mechanisms),
2. whisper and creak (combinable with each other and with category 1), and
3. breathy and harsh voice.

Although these are phonatory settings, such vocal qualities may also be provoked by the physical state of the voice such as fatigue, injury, and illness.

John Esling and co-workers have later specified the contribution to phonation of the false vocal folds and of the constrictor muscles further above, as summarized in [6]. This helps explaining constricted and unconstricted phonation modes and characterizes harsh, creaky and whispery voices as constricted phonation modes and modal voice, falsetto and breathy voice as unconstricted ones.

2.3. Recording of voice qualities

In addition to considerations of the physiology and the acoustics of the voice, we recorded one male and one female actor saying 10 sentences (in French) while faking different voice qualities depending to their abilities. The voice qualities included soft, tense, breathy, hoarse, whispering, nasal and lisping voices, as well as the voice of an old person, a child, a drunk or the effect of a stuffed nose.

Comparison with their normal voice, which was also recorded (at 48 kHz and 24 bits), gave important spectral information for the transformations, as discussed below.

3. SIGNAL TRANSFORMATION

Understanding the physiological mechanisms distinguishing different voices and phonation types is not enough for changing the signal. A model relating the physics of the voice and the emitted signal is also necessary. The optimal model is probably one than separates the glottal source (as far as possible) from the vocal tract. Such source-filter separation algorithms exist [7, 8], but do not yet seem mature for high-quality signal transformation.

Instead, a number of signal-centered methods have been developed, the most successful ones probably being the PSOLA method [9, 10], harmonic plus noise methods (HNM) [11, 12, 13], and the phase vocoder [14, 15, 16, 17, 18]. While all these methods can perform transposition and time stretching of the signal, only the latter two principles allow finer modification of the signal in the frequency domain. An informal listening test showed that our improved version of the phase vocoder (called *SuperVP*) gave more pleasant results than two independent implementations of the HNM [11, 12]. Another reason for choosing the phase vocoder for the transformations presented in this paper is the fact that it is under continuous development for musical applications at Ircam, and serves as the engine of AudioSculpt, a powerful graphically interactive software for music modification [19].

3.1. The phase vocoder and improvements

The basic phase vocoder [14] is roughly a series of band filters, in practice implemented as successive Short-Time Fourier Transforms (STFTs), that reduce the signal into amplitudes and phases in a uniform time-frequency grid. Basic parameters are the window and FFT sizes of the STFT and the time step between each STFT. Combined with resampling and changing of the time step between analysis and synthesis, this method allows for high-fidelity time stretching and pitch transposition as well as modification of the amplitude of each point in the grid, and thus an enormous potential for transformations.

A well-known artifact of the phase vocoder is the introduction of “phasiness”, in particular for speech, the result sounding strangely reverberant or with a lack of presence of the speaker.

Improvements added to our implementation of the phase vocoder and constituting *SuperVP* are: detection and processing of transients [16], waveform preservation for single-source processing [17], robust spectral-envelope estimation [18, 20], and dynamic voicing control based on spectral-peak classification [21].

An informal listening test has been performed, placing SuperVP superior to PSOLA except for downward transposition.

3.2. Basic signal analyses and transformations

While the estimation of the fundamental frequency F0 is not an issue here, F0 is an important component in the transformations, as is a robust decision of whether the signal is voiced or not [22, 23]. Another important property of the speaking voice is the fact that the harmonics in voiced segments of the signal are masked by noise above a certain frequency, which may vary from below 1 kHz to above 4 kHz depending of the voice and the phonatory setting applied. This frequency is called *VUF* in the following. Finally, a robust estimation of the *spectral envelope* is obtained by the cepstrally based true-envelope estimator [20]. Compared to LPC-based methods, it has the advantages of not being biased for harmonic signals and that its order may be adapted automatically to the local F0.

The fact that the true envelope truly follows the spectral peaks of the signal, equips us with the control necessary for detailed filtering in time and frequency depending on the time-frequency characteristics of the signal. As long as the time variation is done with care to avoid audible discontinuities, the results keeps a high degree of naturalness.

When it comes to the basic transformations, by *(pitch) transposition*, we mean changing the local F0 of the signal by a certain factor while conserving the spectral envelope (as far as possible). The transposition of the *spectral envelope* is an independent parameters although both are done in the same operation. Time-frequency filtering has already been mentioned, and a fourth basic transformation time stretching, which does not touch the frequency dimension [14].

Most of these operations work in real time. On the other hand, the voice transformations described in the following have been developed in MATLAB and are not yet available in a real time. They should be, however, by the time that this paper is published.

4. VOICE TRANSFORMATION

We may group transformations of type and nature of the voice into three categories: transformation of physical characteristics of the speaker (size, sex and age), transformation of voice quality (modification of the glottal source to make the voice breathy, whispering, rough, soft, tense, loud etc.), and transformation of speech style (modification of the prosody; liveliness, speech rate, etc.).

4.1. Transformation of size, sex and age

While there are general differences between voices of different sex and age, there are also considerable differences within each category. This makes it difficult to determine absolute parameters for successful transformation of the voice, and even though the parameters would be correct, the perception of sex or age may be disturbed by the fact that the speech style does not correspond to the voice. Nevertheless, with the pitch values given in Sec. 2.2 as reference, modification of pitch to change sex and age may simply be achieved by a transposition of the source signal to the given target pitch.

But merely increasing the pitch of a man's voice does only make a man speak in falsetto, or as Mickey Mouse if the spectral envelope is transposing together with F0 and the harmonics, for instance. The vocal tract should be modified independently by transposing the spectral envelope according to an average of the ratios of the formants of men, women and children given in Sec. 2.2. In order to achieve other voices, such as a teenaged boy or girl, intermediate values were chosen.

In an interactive system, the operator can in addition be given the possibility to optimize the parameters for each voice. In some cases it may indeed be interesting to play with the ambiguity of the sex of the voice and thus choose an intermediate setting, as we did with Céladon's voice when he disguises himself as a woman in the film "Les amours d'Astrée et de Céladon" by E. Rohmer, 2007.

When it comes to aged voices, the characteristics mentioned in Sec. 2.1 are converted into transformations. For the sake of the perceptual evaluation of this paper, a convincing old voice was achieved by the following four measures: (1) Trembling was implemented as a slowly fluctuating transposition factor, (2) the pitch was slightly raised (together with the spectral envelope to give a brighter timbre), (3) the speech rate was slowed down, and (4) the F0 ambitus was decreased. Additionally, breathiness may be added, as described below.

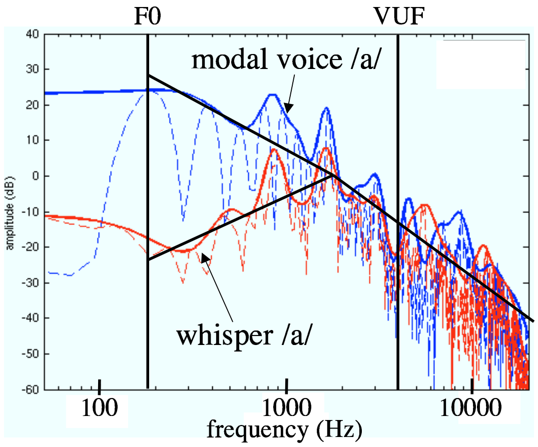

Fig. 1: Spectrograms with spectral envelopes of a voiced and a whispered /a/. VUF is the local voiced/unvoiced frequency.

Finally, no literature was found on transformation of size, but we can simply extrapolate our knowledge about sex and age transformation. The approach is intuitive, as when we read a book aloud for a child: raising the pitch and making the vocal tract smaller (transposing the spectral envelope upwards in frequency) make us sound like a small dwarf, and speaking with a low pitch and making the mouth cavity large (downwards spectral-envelope transposition) simulate the voice of a giant, for instance. Adding breathiness may be an efficient addition to a deep dragon's voice, for instance.

4.2. Whisper

When we whisper, the vocal folds are separated enough not to vibrate but are still sufficiently close to produce audible turbulence. Our recordings showed that the spectral envelope of whisper and voiced speech are comparable at high frequencies (above the estimated VUF) but differ at low frequencies for voiced phones. While the formant frequencies have approximately the same positions, the spectral tilt was flat or even positive below the VUF, as seen in Fig. 1.

To transform voiced speech to whisper, a source of white noise was therefore filtered by the spectral envelope estimated from the original signal, except for some modification at low frequencies: Firstly, the spectral tilt was neutralized and even inverted below about 3 kHz, depending of the voice. The choice of 3 kHz was an empiric compromise because using the VUF as cut-off frequency for the inversion tended to create audible discontinuities.

Secondly, *fricatives* (unvoiced phones such as /f/, /s/, /θ/, /ʃ/) should not be touched by this transformation as they depend on turbulence created at constriction further downstream. Since these noisy sounds have the energy concentrated at higher frequencies (in the range 3–6 kHz depending on the sound [24]), preserving the fricatives was indirectly achieved by the measure described above by allowing only to increase the low-frequency spectral tilt.

4.3. Breathy voice

A breathy phonation is obtained by reducing the force with which the vocal folds are pressed together (by slightly opening up the glottis or due to a fatigue). The vocal folds vibrate, but the closing movement is not complete or sufficiently soft for air leakage to cause turbulence noise in addition to harmonics. The effect of this on the spectrum is an increasing spectral tilt of the harmonic parts of the signal (i.e., an attenuation of high-frequency harmonics) accompanied by an addition of aspiration noise above about 2 kHz [25].

To render a voice breathy, we proceed in to steps: first, the voice must be *softened* by passing it through a lowpass filter, then noise must be added. Of course, the noise must change with the signal, which is exactly the case for the spectral envelope. An approach similar to that of whisper is thus followed, and the noise is attenuated at low frequencies to avoid it to interfere with the original signal. However, just as with whisper, the fricatives should not be touched. It is therefore important to modulate the lowpass filter with the voicing coefficient. The original, lowpass-filtered signal is then mixed with the whisperlike noise at a ratio that depends on the desired degree of breathiness.

4.4. Transformation of speech style

Differences between sex and age are also seen in terms of speech style. Speech style is much more difficult to address from a global point of view because it necessitates at prosodic analysis and processing. It was recognized, however, that the dynamic range of the pitch, the *pitch ambitus*, is a speech-style attribute which varies from speaker to speaker and seems in general greater for children and teenagers than for adults, and even smaller for aged people.

Changing the ambitus was efficiently achieved by exaggerating or attenuating the natural variations of F0 by dynamically transposing the signal in proportion to the log-F0 deviation from the established median F0. The median F0 was chosen rather than the mean F0 because the median is invariant of the method used for this transformation.

Another speech-style attribute is the *speech rate*. Slowing down the speech to get an aged person, for instance, was done by dilating the signal by some 20 to 50 % without changing the pitch or spectral envelope.

Changing the ambitus and the speech rate together has surprising effects: *dullness* may well be achieved from a neutral recording by decreasing the speech rate and the ambitus. Conversely, the opposite transformation of the same neutral recording gives the effect of *eagerness*.

5. PERCEPTUAL EVALUATION

Two successive listening tests were realized in order to assess the naturalness of the transformed voices at three levels: whether the target sex and age were attained, the humanlikeness of the voice, and the quality of the sound. The first listening test was based on transformations piloted by default parameters while the second test was held after manually optimizing the parameters to the individual characteristics of the source voices.

5.1. Stimuli and test setup

The stimuli were based on 13 French-speaking voices, of which 5 women, 6 men, 1 girl, and 1 boy, two sentences (none the same) spoken by each voice, in total 26 recordings of about 2 to 3 seconds. Except for the girl speaking spontaneously, all were reciting a text. The quality was at least studio quality, except for one woman, one man, and the girl, and recorded at sampling rates ranging from 16 to 48 kHz.

For each of the 26 sentences, 8 variants were used: 2 sexes × 4 ages (1=child, 2=teenager, 3=adult, and 4=old person), i.e., 7 transformations and the unmodified original. The order of the voices was randomized, and no subject was to hear the same sentence twice, nor the same transformation on the same voice.

While the idea was to compare the results of the two tests, the first one provoked a slight redesign of the test based on comments from the listeners. Firstly, the listening test was established as an Internet page written in PHP. The great advantage is that the subjects may themselves choose whether, when, and where to pass the test, but the disadvantage is that the listening conditions could not be fully controlled. To compensate for the lack of control, we asked the subjects to do the test in a silent environment and asked them what kind of sound source they used (headphones, ear-plugs, or PC or HiFi loudspeakers) as well as their level of understanding of the French language (as native, well, or not well). The audition of the subjects was not evaluated.

In the first test, the list of the 26 stimuli presented for the subject was randomized on the fly by the PHP server, but for some reason, the female adult case was never presented. In the second test, an even distribution of source voices and transformation variants was assured by preparing 8 random sets of stimuli covering all the stimuli once and only once. Futhermore, the design was improved somewhat, so we concentrate this paper on the second test.

For each stimulus, four questions were presented on the screen. The subjects were forced to guess the sex (with a possible doubt) and age (the same four ages as above in addition to a "too ambiguous" choice). The third question was whether the voice sounded like a human with three choices (Yes, a human speaking naturally; Yes, a human speaking in an unnatural way; or No, not human. Finally, we asked the subjects to evaluate on a 5-point DMOS scale whether they noticed any artifacts (No; Yes, but not annoying; Yes, slightly annoying; Yes, annoying; and Yes, very annoying). The term artifact was described as an extraneous sound, noise or effect such as "buzzing", echo, doubled voice, or a metallic or robotic character accompanying the sound, whether for a brief moment or persisting throughout the recording.

Before starting the test, the subjects were informed of the purpose of the project and that transformed samples and authentic recordings were randomly interchanged. Three examples representing extremes were given, accompanied by proposed judgements: natural human voices with no audible artifacts and with very annoying and non-human voice (like Mickey Mouse) with slightly annoying artifacts.

In the first test, there was no choice for ambiguous age, but a fourth choice to the humanlikeness (before "not human"): "caricature of a human", without the words Yes and No. Furthermore, the last question was "How do you consider the sound quality?" supported by a 5-point MOS scale from excellent to bad.

5.2. Results

Of the 31 subjects participating in the second listening test, all but 2 had French as mother tongue (27) or understood French well (2). 20 used headphones, 3 used ear-plugs, 6 used PC loudspeakers or similar, and 2 used HiFi loudspeakers. All judgements were therefore pooled together in the following analysis.

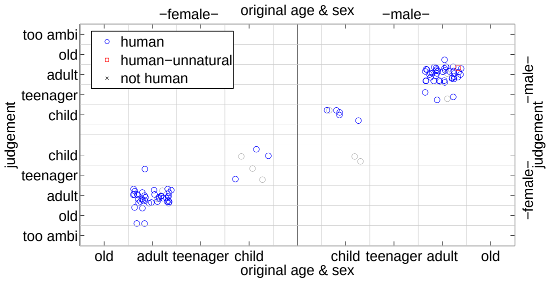

Fig. 2: Correspondence of intended (horizontal) and judged (vertical) sex and age for the unmodified samples. Colors/shapes judgement of naturalness, grey samples signal uncertain sex.

Figure 2 shows judged sex and age as function of the voices' actual sexes and ages. Male voices are at the right and upper part of the graph while female voices to the left and below. If judged and actual sex and age are equal, the points should lie along the oblique line $y = x$. This is mostly the case for the originals.

The humanlikeness of the voice is indicated with shapes (and colors): (blue) circles mean humanlike, (red) squares have been judged as human speaking unnaturally, and (black) crosses indicate non-human voices. Some markers are grey, which signal that the subject judged sex with hesitation. Samples that were judged to have ambiguous age are gather in the upper- and lower-most lines of the plot.

The transformed voices are plotted in the same way in Fig. 3 with the *intended* sex and age along the horizontal axis. The points are fairly well concentrated around the oblique $y = x$ line, but there are a quite a few "wrong" judgements. In particular are the voices intended to be female mostly evaluated as female. The most striking exceptions are that the age of a number of old and girls' voices was considered ambiguous, and teenager voices being confused with adult ones. Concerning the voices intended to be male, we see that boy voices were often confused with a woman's voice, teenagers with adults, and some old men's voices were merely considered that of a man. However, since the sex and age of the source voice do not appear in this representation, reasons for the errors are difficult to discuss.

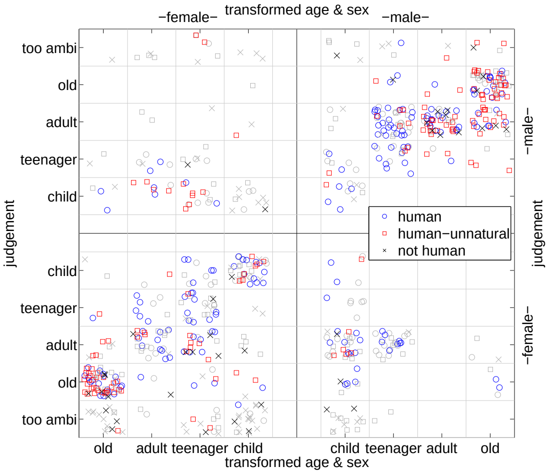

Fig. 3: Correspondence of intended (horizontal) and judged (vertical) sex and age for the transformed samples (excluding the originals). Colors/shapes judgement of naturalness, grey samples signal uncertain sex.

In the following we therefore try to quantify the success of the transformations depending on the source voice. The judgements were therefore grouped depending on source voice and intended sex and age and counted according to the recognition of sex and age and the scores on humanlikeness and sound quality:

1. sex and age were recognized as the intended sex and age *and* voice was judged as humanlike (human or human speaking unnaturally),
2. artifacts were audible and slightly annoying or better, and
3. artifacts were audible but not annoying or better.

Uncertain sex was counted while ambiguous age was considered not recognized. Table 1 shows the percentages of subjects that recognized sex and age of the voice and judged the voice humanlike (group 1). The first 6 source voices (m3) were male adult voice, the next 5 (f3) female adult voices, then the boy (m1) and finally the girl (f1). The number of judgements for each group is shown after the slash. The last two lines contain the averages for the male and female adult voices with standard deviation in parentheses. A great variability depending on source voice and on transformation target is seen. The mean percentage for transformed recordings based on men's voices was 29±28 % and 43±32 % for transformed women's voices (87±16 % for their originals independently of the sex of the original). It is fairly clear from these results that the women's voices were easier to transform to another sex and age than men's voices. Setting the threshold for acceptance at the majority (i.e., more than 50 %, highlighted with bold type face), the *women's* voices showed to transform well to *old woman* and *girl* as well as *man* and *old man*. The men's voices were less flexible and mostly useful only to make old women and old men. Astonishingly, one man's voice (M) was less ambiguous when transformed to a woman than as original. There was nothing wrong with his voice identity, but a closer examination showed that one subject had judged his voice as old. This demonstrates that the numbers are prone to rather large errors due to the fact that there were few judgements per voice. For practical means, it may be said that half of the men's voices gave 1 transformation accepted by the majority, the other half 2 new voices. For the woman's voices 2, 3, or 4 new voices could be made, depending on the source voice. The boy's voice was accepted in 3 other variants, while the girl's voice did not transform convincingly, following the majority of the subjects.

Similar tables were set up for the percentage of subjects that judged artifacts slightly annoying or better (group 2) and not annoying and better (group 3). The original recordings were in general considered to have at worst not annoying artifacts by a great majority of the subjects. A few of the transformed recordings could compare to these, but as for the recognition of sex and age, there were great differences between the different source voices dependent of the transformation intended. On average, 69±20 % of the subjects considered that the modified men's recordings had slightly annoying artifacts or better, 79±18 % for the women's voices. Increasing the requirement to not annoying artifacts decreases the averages to 47±21 % and 58±18 % respectively. As for the recognition of sex and age, transformation of the women's voices scored higher on sound quality than for men's voices, especially when taking into account that the original recording of one of the women's voices (E) was judged to have worse than slightly annoying artifacts by 43 % of the subjects.

5.3. Discussion of test setup

A number of choices were made in the design of the listening test. The humanlikeness scale was an experimental scale. In the first listening test, it had four alternatives: human, human changing his voice, caricature of human, and non-human. The

Table 1: The percentage of subjects recognizing the sex and age *and* judging the voice as humanlike, in bold if more than 50 %. The number of judgements are shown after '/', and the average percentages for men (m3) and women (f3) are listed at the end with standard deviation in parentheses

		female				male			
voice	orig	old	adult	teenager	child	child	teenager	adult	old
P (m3)	**100**/7	**57**/7	44/9	0/8	0/7	0/9	0/9	–	**100**/6
J (m3)	50/6	29/7	0/8	0/6	27/11	25/8	**60**/10	–	17/6
R (m3)	**100**/8	29/7	0/7	29/7	38/8	43/7	11/9	–	**67**/9
C (m3)	**100**/9	**62**/8	50/6	11/9	0/7	0/9	14/7	–	29/7
M (m3)	**86**/7	33/9	**100**/8	0/9	29/7	0/7	0/6	–	**89**/9
X (m3)	**88**/8	**60**/10	**64**/11	33/6	33/6	17/6	0/7	–	29/7
D (f3)	**100**/9	**83**/6	–	44/9	43/7	0/9	0/7	**57**/7	25/8
E (f3)	**86**/7	**88**/8	–	33/9	**71**/7	0/7	17/6	**56**/9	44/9
A (f3)	**100**/9	44/9	–	29/7	**83**/6	14/7	12/8	**100**/9	43/7
F (f3)	**75**/8	**100**/11	–	33/6	0/6	0/6	0/8	**71**/7	**80**/10
O (f3)	**78**/9	**57**/7	–	43/7	**67**/9	0/8	14/7	**71**/7	**88**/8
H (m1)	**75**/8	33/6	33/6	0/10	**75**/8	–	50/6	**86**/7	**73**/11
G (f1)	50/6	0/7	38/8	33/6	–	25/8	40/10	45/11	17/6
man	**87** (19)	45 (16)	43 (39)	12 (15)	21 (17)	14 (18)	14 (23)	–	**55** (35)
woman	**88** (12)	**74** (23)	–	37 (7)	**53** (33)	3 (6)	9 (8)	**71** (18)	**56** (27)

idea was to have an intuitive scale. Nevertheless, some subjects complained that in particular the caricature alternative was difficult to understand. In the second test, the caricature alternative was therefore removed, and "yes" and "no" were added in order to make clear that the middle alternative was still human while keeping the alternatives intuitive. However, both these scales made it impossible to calculate the average of several judgements without deciding how to weight the middle alternative(s). Probably, a 5-point symmetric scale would be the best although less intuitive:

Yes, certainly human (1);
Yes, probably human (0.5);
Uncertain (0);
No, probably non-human (-0.5); and
No, certainly non-human (-1).

For the sound-quality scale, suggestions existed already: the ITU-T recommendation P.800 presents a number of scales, and in particular the mean opinion score (MOS) scale (excellent, good, fair, poor, and bad), and the degradation MOS (DMOS) scale (degradation is inaudible, audible but not annoying, slightly annoying, annoying, and very annoying). The first test placed the sound quality on the MOS scale, but again, certain subjects did not find "sound quality" sufficiently precise. This is why we adopted a DMOS scale and asked whether artifacts such as "buzz, echo, strange sounds/noises, etc." were noticeable and annoying.

Of course, these adaptations made a direct comparison between automatically parameterized transformations (test 1) and manually optimized ones (test 2) difficult. This is also why we have offered the first test little attention. However, it is clear that the judgements of humanlikeness and recognition of sex and age were better in the second test, which implies that the default parameters may be improved, although this may require a more fundamental model for the analysis of the source voice.

5.4. Discussion of results

It was disappointing that the measure of success of the transformation of sex and age indicated that, depending on the source voice, only 1 to 4 convincing new voices (except 0 for the girl) could be made from a given source voice. However, in a number of cases, it was also observed that some transformed voices were more or less consistently classified as a different voice than intended, which is not taken into account in the table. The example of the original of voice M being wrongly classified is a good example. This confirms that judgement of sex and age is a subjective one, and maybe in particular for transformed voices.

It may also be argued that using more time to optimize the parameters, may improve the scores of the listening test. But this does not change the fact that much information about the sex and age of a speaker is conveyed not only in the pitch and timbre of the voice, but also in the prosody, vocabulary, and the sentence structure. This becomes very clear with the recording of the girl, whose transformations experienced the worst recognition of all voices. Transforming her voice into a deep man's voice makes a great contrast to the childish way she spoke. The boy, however, whose speech was controlled by the written text, was rather well judged for many transformations.

While the scores for sex and age recognition are connected to the sex- and age-transformation parameters, the sound quality judgements should be related the signal-processing method. For instance, a woman's voice has a start pitch and a spectral envelope that are intermediate to the other voice types while transforming a man's voice to a child's voice requires a transposition of pitch approaching two octaves. It is well known that the greater the modification of the signal is, the more the naturalness of result is disturbed. This relation is directly reflected by decreasing percentages for increasing acoustic distance between source and target voice.

The requirement of sound quality depends on context. Furthermore, it should be noted that in the setting of this kind of listening tests, the subjects are keen on finding artifacts. In real applications, however, the attention of the audience is also occupied with visual impressions and the story line, for instance. In addition, artifacts may also be masked by background sounds such as music and traffic noise etc. So while high-fidelity applications may seem to require artifacts to be not *audible*, our experience shows that artifacts that are audible but not annoying or even slightly annoying in a silent context may not be *noticeable* in the actual context.

6. CONCLUSIONS

Methods for transformation of sex and age, voice qualities whisper and breathy, and speech style have been presented.

The perceptual evaluation concerning sex and age transformation showed that

(1) transformed voices scored worse than original recordings with all respects except in a few cases,

(2) the success of a transformation depends to a great extent on the source voice,

(3) 1 to 4 transformations out of 7 (2 sexes × 4 ages except the original) were considered human-like and correctly classified in terms of sex and age by the majority of subjects, women's voices showed to transform more easily than men's voices and give one successful transformation more than them, and speech-style transformation should be considered for sex and age transformation,

(4) the transformation of sex and age of women's voices gave less artifacts than the transformation of men's voices, and

(5) a more fundamental model for the analysis of the source voice seems necessary for better automatic transformation of the sex and age.

A supplementary listening test should be performed in order to evaluate the naturalness of the other transformations.

Sound examples are available at
http://recherche.ircam.fr/anasyn/farner/pub/AES09.

ACKNOWLEDGEMENTS

The work was done as a part of the projects VIVOS and Affective Avatars with financial support from Agence Nationale de la Recherche (ANR). We are grateful to Chinkel (SonicVille) for recording of actors, and not least the two actors Christine Paris and Matthieu Rivolier for their engaging interpretations of a great number of voice qualities. The two anonymous reviewers are also thanked for helpful comments to the precis.

7. REFERENCES

[1] Gordon E. Peterson and Harold L. Barney, "Control methods used in a study of the vowels," *Journal of the Acoustical Society of America*, vol. 24, no. 2, pp. 175–184, 1952.

[2] Ke Wu and D.G. Childers, "Gender recognition from speech. Part I: Coarse analysis," *Journal of the Acoustical Society of America*, vol. 90, no. 4, pp. 1828–1840, 1991.

[3] Markus Iseli, Yen-Liang Shue, and Abeer Alwan, "Age, sex, and vowel dependencies of acoustic measures related to the voice source," *Journal of the Acoustical Society of America*, vol. 121, no. 4, pp. 2283–2295, 2007.

[4] R. J. Baken, "The aged voice: A new hypothesis," *Journal of Voice*, vol. 19, no. 3, pp. 317–325, 2005.

[5] John Laver, *The phonetic description of voice quality*, Cambridge studies in linguistics. Cambridge University Press, 1980.

[6] Lisa Danielle Bettany, “Range exploration of phonation and pitch in the first six months of life,” Master of arts, University of Victoria, 2002.

[7] W. Ding, H. Kasuya, and S. Adachi, “Simultaneous estimation of vocal tract and voice source parameters based on an arx model,” *IEICE Trans. Inf. Syst.*, vol. E78-D, no. 6, pp. 738–743, jun 1995.

[8] Damien Vincent, Olivier Rosec, and Thierry Chonavel, “A new method for speech synthesis and transformation based on an ARX-LF source-filter decomposition and HNM modeling,” in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[9] Eric Moulines and Jean Laroche, “Non-parametric techniques for pitch-scale and time-scale modification of speech,” *Speech Communication*, vol. 16, pp. 175–205, 1995.

[10] Geoffroy Peeters, *Modèles et modification du signal sonore adaptés à ses caractéristiques locales*, Ph.d. thesis, Ircam, Paris, France, 2001.

[11] Yannis Stylianou, *Modèles Harmoniques plus Bruit combinés avec des Méthodes Statistiques, pour la Modification de la Parole du Locuteur*, Ph.d. thesis, Ecole Nationale Supérieur des Télécommunications, Paris, France, 1996, (in French).

[12] Marine Campedel-Oudot, *Application du modèle sinusoides et bruit au codage, débruitage et à la modification des sons de parole*, Ph.d. thesis, Ecole Nationale Supérieur des Télécommunications, 1998, (in French).

[13] Xavier Serra and Julius Smith III, “Spectral modeling synthesis: A sound analysislsynthesis system based on a deterministic plus stochastic decomposition,” *Computer Music Journal*, vol. 14, pp. 12–24, 1990.

[14] M. Dolson, “The phase vocoder: A tutorial,” *Computer Music Journal*, vol. 10, no. 4, pp. 14–27, 1986.

[15] Jean Laroche and Mark Dolson, “Improved phase vocoder time-scale modification of audio,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, 1999.

[16] Axel Röbel, “A new approach to transient processing in the phase vocoder,” in *ProcİInt. Conference on Digital Audio Effects (DAFx-03)*, London, UK, Sept. 2003, pp. 344–349.

[17] Jean Laroche, “Frequency-domain techniques for high-quality voice modification,” in *Int. Conf. on Digital Audio Effects (DAFx) 03, London, UK*, 2003.

[18] A. Röbel and X. Rodet, “Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation,” in *Proc. Int. Conference on Digital Audio Effects (DAFx-05)*, Madrid, Spain, 2005, pp. 30–35.

[19] Niels Bogaards and Axel Röbel, “An interface for analysis-driven sound processing,” in *119th AES Convention*, oct 2005.

[20] Axel Röbel, Fernando Villavicencio, and Xavier Rodet, “On cepstral and all-pole based spectral envelope modeling with unknown model order,” *Pattern Recognition Letters*, vol. 28, pp. 1343–1350, 2007.

[21] Miroslav Zivanovic, Axel Röbel, and Xavier Rodet, “Adaptive threshold determination for spectral peak classification,” *Computer Music Journal*, vol. 32, no. 2, pp. 57–67, 2008.

[22] Alain de Cheveigné and Hideki Kawahara, “Yin, a fundamental frequency estimator for speech and music,” *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, apr 2002.

[23] Chunghsin Yeh, *Multiple fundamental frequency estimation of polyphonic recordings*, Ph.D. thesis, University Paris 6, France, 2008.

[24] Martin F. Schwartz, “Identification of speaker sex from isolated, voiceless fricatives,” *Journal of the Acoustical Society of America*, vol. 43, no. 5, pp. 1178–1179, 1968.

[25] Dennis H. Klatt and Laura C. Klatt, “Analysis, synthesis, and perception of voice quality variations among female and male talkers,” *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, feb 1990.