The OMe (Octave-Median) scale: a natural scale for speech melody.

Céline De Looze[1], *Daniel Hirst*[2,3]

[1]Speech and Phonetics Laboratory, CLCS, Trinity College, Dublin, Ireland
[2]Laboratoire Parole et Langage, CNRS & Aix-Marseille Univ., France;
[3]School of Foreign Languages, Tongji University, Shanghai, China

deloozec@tcd.ie, daniel.hirst@lpl-aix.fr

Abstract

Fundamental frequency, the primary acoustic correlate of speech melody, is generally analysed and displayed using a linear scale (Hertz) or a logarithmic one, generally in semitones and usually offset to an arbitrary reference level such as 100 Hz. In this paper we argue that a more natural scale for analysing speech is the OME (Octave-MEdian) scale, using the octave (o) as the basic unit, offset to the median value of the speaker's range. We present results showing that a reasonable estimate of a speaker's neutral pitch range can be obtained directly from the median.

Index Terms: Tone, intonation, melody, pitch, octave.

1. Introduction

Although one can observe some non-linearity in the perception of the pitch of speech sounds, fundamental frequency is unquestionably the main acoustic correlate of perceived pitch height. Psycho-acoustic scales for the study of speech have also been proposed, particularly the *Mel*, *Bark* and *ERB* scales. The relevance of these scales, however, remains unproven. A recent study [25], for example, using a task of replicating pitch contours between male and female voices, showed that a logarithmic scale better reflects the performance of speakers than either a linear or a psycho-acoustic scale. The physical scale in *Hertz* (cycles per second) is very often transformed, in studies of prosody, to a logarithmic scale, most often expressed in semitones with a reference value (called C0), arbitrarily set at 16.3516 Hz [28]. This reference was chosen as being close to the lowest pitch perceptible to the average human ear and as well as corresponding to the tuning of A_4 (= 'A above middle C') to 440 Hz, in conformity with the ISO Standard ISO16 [19].

Fant [12] proposed the St (= semitone) unit defined as:

$$St = 12 \cdot \frac{ln(\frac{Hz}{100})}{ln(2)} \qquad (1)$$

The semitone is, however, in no sense a natural unit of measurement. It is, in fact, the product of a complex history of Western classical music culture, corresponding to the division of the octave into 12 equal intervals, an idea that had first been described in a treatise published in China in 1584 [20]. In Europe, the scale of 12 equal semitones equal (= *equal temperament*), has been used increasingly since the 18th century to tune keyboards, replacing the *natural scale* ('just intonation') previously used, or Bach's *well-tempered scale*. All these scales were the result of a search for a compromise which would allow musicians to modulate from one scale to another without introducing major discord and without having to switch keyboards.

In different civilizations at different times, the use of different sets of notes can be observed. Practically all of these scales, however, have in common the fact that the names of the notes are generally the same, regardless of the octave. Thus, in the Western classical scale, for example, the sequence *Do Re Mi Fa Sol La Ti Do Re Mi ...* etc. can be repeated indefinitely within the physical limits of sound production.

This circularity (also known as *chromatic repetition*) seems to stem from a physiological basis of human perception [6, 5] including that of neonates [22] and also that of rhesus monkeys [27]. It was observed as early as the 60s, in an anatomical study of a cat, that the auditory thalamus is organised in stacked layers or laminae. It was suggested that this organisation may have a specific function in the processing of acoustic frequencies [24]. [23] and [18] later demonstrated that the auditory thalamus of the cat actually contains a neural chroma map, underlying an octave architecture. While the functional role of the mammalian auditory thalamus octave topography still needs to be determined, recent research has suggested that it may cause, as a side effect, the octave circularity of pitch that has been observed in the rhesus monkey as well as in humans [6]. This study investigated the effect on a musician with absolute pitch, of the neurotropic medical drug carbamazepine (CBZ), known to have a down-shift pitch side effect,in order to better understand the mechanism of octave circularity of pitch. They observed in their subject, during a pitch identification task, an internal tone-scale or chroma representation. When CBZ was taken, pitch shift was indeed observed but the pattern of tone representation remained unchanged. This suggests that the human brain may be hard-wired for octave-circular pitch perception.

In any case, it is the octave, not the semi-tone, which appears clearly as the basic unit for the natural perception of the pitch of speech sounds and music.

The use of the *semi-tone* (or in more precise studies, its subdivision the *cent* [where $1st = 100cents$) has paradoxically had the negative effect of masking the importance of the octave as a basic unit in pitch production and perception. Rereading a number of studies on pitch range with this in mind reveals a very large number of cases where authors report an interval close to an octave (= 12sts) or half-octave (= 6sts) without drawing attention to this fact, or perhaps, even, sometimes without having noticed it. In [26], for instance, the authors reported a f0 mean at the beginning of sentences produced in neutral, happy, angry and scared voices of 6.72, 12.64, 12.52 and 12.38 sts respectively. If we calculate the difference between the neutral voice mean f0 and the other voices mean f0, it reveals for each 'arousal' voice a shift of half-an-octave.

10.21437/SpeechProsody.2014-170

The intervals octave and half-octave may play a specific role in speech production. [4] investigated the pitch contours of utterances produced under two conditions (in a normal voice in a quiet room vs. in a louder voice when exposed to noise over headphones), and observed for instance a raising of half-an-octave for the increased loudness condition. A raise of a half-an-octave or an octave may be used to convey specific linguistic and paralinguistic functions in speech, e.g. signaling focus, topic change, turn-taking as well as expressing arousal.

2. The octave as the basic unit for the perception of pitch.

We recommend the systematic use of the octave (o) and its subdivision the millioctave (mo) for the study of pitch. For precise measurements, the (mo) gives, in fact, approximately the same degree of precision as the cent $[1mo = 1.2cents]$ and has the advantage of being in conformity with the general practice of the *International System of Units: SI,* in which prefixes corresponding to an exponent divisible by 3 are generally preferred.

As a derived *SI* unit, the octave can be defined as:

$$o = log_2(s^{-1}) \tag{2}$$

where s is the duration in seconds of a period.

The second author has suggested elsewhere [13, 14] that there may also be a physiological explanation for the octave and half octave as a basis for the production of melodic intervals. [14] reported an experiment where these two intervals were observed as modal values in a task of producing varied contours on isolated syllables in French, *oui* and *non*.

In so far as the vocal folds behave like vibrating strings, the relationship between tension and frequency is governed by Mersenne's law which states that

> *The frequency of a vibrating string is proportional to the square root of its tension.*

A doubling of the tension would consequently correspond to a rise of half an octave. This might explain why the intervals - octave and half octave - seem to be frequent in the production of speech melody, even though a rise or fall of an octave on a single syllable is certainly not perceived in its entirety.

3. Corpora

Four corpora, a total of about 2 hours of speech, were selected for this study: these consisted of extracts of the *PFC* and the *CID* corpora for French and the *PCA* and *Aix-MARSEC* corpora for English. These are briefly described below.

The PFC corpus (Phonology of Contemporary French) [9].

We selected 10 French speakers, from the region of Marseille, 6 female and 4 male. We chose the recordings of their production reading aloud a passage of text, such as an extract from a regional newspaper. This corresponded to approximately 30 minutes of recording.

The CID corpus (Corpus of Interactional Data) [2].

We selected six French speakers, from the region of Marseille, 3 male and 3 female. The recordings corresponded to conversations, where speakers discuss either professional conflicts or unusual situations in which they had found themselves, a total of 30 minutes of recording.

The PCE corpus (Phonology of Contemporary English) [7].

We selected eight English speakers, four male and four female, from the North of England. We chose the the recordings of their production reading aloud a passage, such as an extract from a regional newspaper. This corresponded to approximately 25 minutes of recording.

The Aix-MARSEC corpus [1]. The recordings correspond essentially to extracts from the BBC made in the 1980's.

We selected 51 speakers, 13 women and 38 men. 11 types of production are represented: comments, newsletters, public speech, religious programs, documentaries, fiction, poetry, dialogues, propaganda, etc. This represented a total of approximately 50 minutes of recording.

4. Estimation of pitch range

In [11] we reported results based on the 4 corpora, in English and French, described in the previous section, which showed that, in the production of natural speech, the lower range of fundamental frequency systematically corresponds to half an octave below the median pitch of a speaker's voice, and the upper range generally extends between half an octave and one octave above the median.

The pitch range of a speaker (ie the tonal space actually used in an utterance) is generally measured by two parameters: its *height* (or key) and its *extent* (or span) [21]. Pitch height is generally measured by taking the mean or the median of the distribution of f_0, or by taking the mean value of points considered as representative targets. The span of the pitch range can be calculated by comparing the minimum and the maximum value produced in an utterance or the average values for high and low targets.

A measurement of pitch based on the analysis of tonal targets can be both costly and error prone, especially if the targets are annotated manually.

In this study, we calculated the value of pitch range with respect to the median of the distribution of f0 since this is more stable than the mean which is influenced by extreme values some of which may be erroneous. In addition, the median is a non-parametric measure, independent of the unit or scale of measurement. The median value, in other words, is always the same value whether it is measured on a linear scale or on a logarithmic scale or on one of the psyco-acoustic scales mentioned above.

To avoid the problems inherent in manual measurements, our calculations were carried out using the *Momel* and *INTSINT* algorithms [15].

The Momel algorithm takes as input the raw fundamental frequency curve and models it as the sum of two components, a 'macroprosodic' component on the one hand, consisting of a smooth continuous underlying function corresponding to the intonation pattern of the utterance, and a 'microprosodic' component consisting of a sequence of functions, some of which are discontinuous and which correspond to the local effect of the different individual speech sounds. For more details on these two components cf [16], for a recent attempt to model the 'microprosodic' component for speech synthesis in Arabic, cf [8].

The INTSINT algorithm takes as input the target points detected by the automatic pitch modeling algorithm (Momel) and codes these targets using an alphabet of 8 discrete symbols. The symbols **T**(op) and **B**(ottom) delineate *high* and *low* values of the speaker's pitch range, respectively, while **M**(id) codes its central tendency. Targets coded **H**(igher), **L**(ower), **S**(ame),

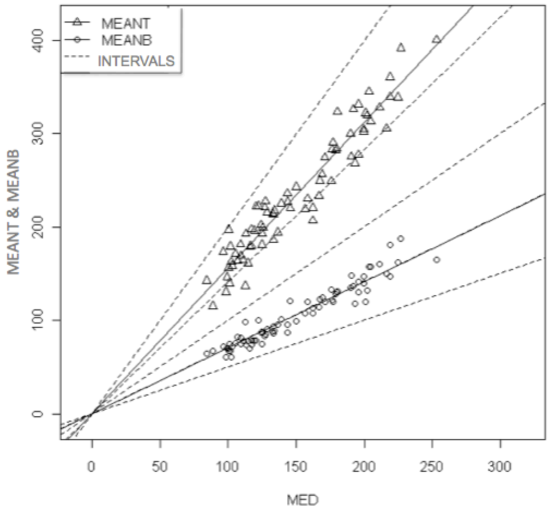

Figure 1: Linear regressions corresponding to Mean-**B** and Mean-**T** are traced in continuous lines and the dotted lines represent, from top to bottom, the intervals +octave, +half-octave, unison, -half-octave and octave compared to the median.

U(pstepped) or **D**(ownstepped) are defined not globally but locally, with respect to the value of the immediately preceding pitch target and are defined as being, respectively, *higher*, *lower*, *equal to*, *slightly higher* or *slightly lower* than the preceding target. For the precise definitions of the coding cf. [15].

In this study, we focus on the values obtained for the absolute **T** and **B** tones since all the other values are dependent on the value of the preceding target.

As we mentioned, one of the most common ways to measure a speaker's pitch range is by comparing the mean values of pitch which have been identified as corresponding to high tones and low tones. It is instructive to look at the correlations between the mean values of the low tones (**B**) and high tones (**T**), as determined by the INTSINT algorithm, and that of the *median* of the pitch distribution.

We find, in fact, two strong correlations. Affine relations are as follows:

$$B = 0.741 * median - 5.52$$

$$T = 1.537 * median + 3.75 \tag{3}$$

Significance tests of regression coefficients are highly significant $p < 2^e - 16$. The critical probabilities of the offsets are, however, not significant (p = 0.161 and 0.659). An adjustment of the model without the offset gives:

$$B = 0.706 * median$$

$$T = 1.561 * median \tag{4}$$

For **B**, we find a coefficient of determination (R^2) of 0.92 and for **T** 0.91. This means that it is possible, at least as a reasonable approximation, to predict the limits of the register of a speaker and hence its span, from the median of the distribution of f_0.

ANOVAs on the prediction of average low (**B**) and high (**T**) tones from the median showed no effect of either sex (p = 0.0917 (**B**) and 0.381 (**T**)) or language (p = 0.170 (**B**) and 0.274 (**T**)), or type of production (p = 0.134 (**B**) and 0.368 (**T**) on the slopes of linear regression. It is therefore possible, from the value of the *median*, regardless of the sex of the speaker, both in English and French and whatever the style of speech, to make a reasonably good prediction of the limits (span) of the pitch range. In fact, [10] showed that the relationship between the height and span of the pitch range is actually more complex. In the model given in 4, the range in Hz is strictly proportional to the height because the relationship between T and B is fixed. An even better correlation is obtained with the values on a logarithmic scale, corresponding to a model where the span in octaves is proportional to the height, going from one octave for a low-pitched voice to a little less than an octave and a half for a high-pitched voice.

This co-variation was pointed out by [21] in his discussion of pitch range. The author explains that the difficulty of admitting two dimensions for the register is that these two dimensions co-vary.

5. The OMe scale. A natural scale for the melody of speech

It is interesting to note in the relationships defined in (2) that the coefficient 0.706 corresponds almost exactly to half an octave($log2(0.706) = -0.502$) and the coefficient 1.561 is just slightly over half an octave ($log2(1.561) = 0,642$). We can therefore conclude that the average of the high tones and the average low tones, i.e. the limits of the range of a speaker, for unemphatic speech, usually correspond to about an octave centered on the speaker's median.

This led us to propose [11] a new scale of measurement: the *OMe* (Octave-MEdian) scale defined by the formula:

$$ome = log_2(\frac{Hz}{median}) \tag{5}$$

where *median* corresponds to the median value of f_0 for the recording.

Figure (1) gives a graphical representation of the average low tones (**B**) and the average high tones (**T**) compared to the speaker's *median* pitch. The corresponding linear regressions are plotted in solid lines and dotted lines represent the intervals + *octave*, + *half-octave*, *unison*, - *half-octave* and - *octave* with respect to the *median.* The linear regression on the mean of the low tones (**B**) coincides with the half-octave below the median so that the two lines are not distilnguishable in the figure. That of the average of the high tones (**T**) falls between half an octave and one octave above the median. These musical intervals, defined relative to the median, can therefore be used to estimate the range of a speaker with a reasonable reliability.

They also allow us to propose, as suggested above, a natural scale for the analysis and visualisation of the melody of speech defined in octaves, centered on the median, which we call the *OMe* (Octave-Median) scale.

Figure 2 shows an example of expressive speech by a radio broadcaster pronouncing the sentence 'He dra**MA**tically flourished a **CO**py of Time from nineteen fifty-**THREE**.". As can be seen, the places where the pitch goes above half an octave above the median, correspond precisely to the parts of the words which are perceived as being emphasised expressively.

Figure (3) illustrates the sentence "What can I have for dinner tonight? " read by one male and one female speaker.

The visualisation of these recordings was obtained automatically from the signal and TextGrid using the Praat plugin

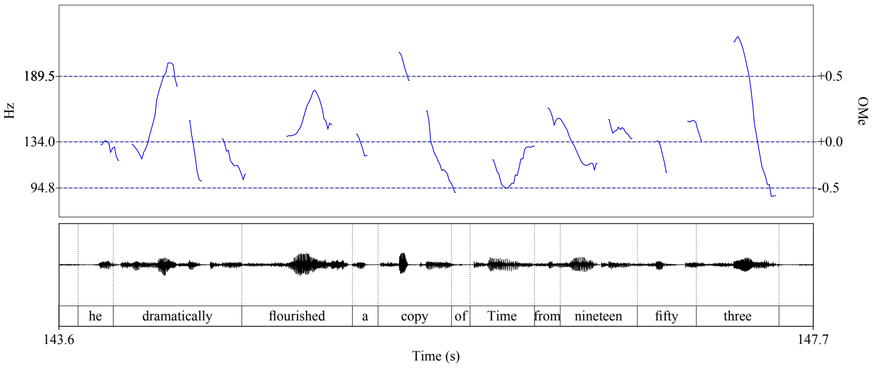

Figure 2: An extract of journalistic speech “He dra**MA**tically flourished a **CO**py of Time from nineteen fifty-**THREE**.”

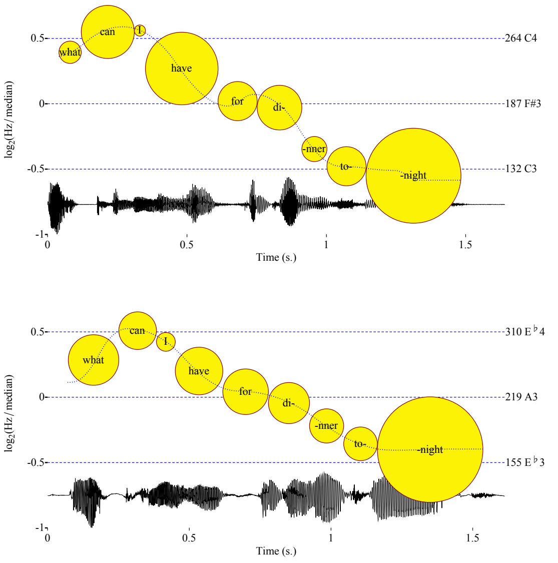

Figure 3: Graphical representation of readings of the sentence ”What can I have for dinner tonight.” by one male (top) and one female (bottom)speaker, displayed using the *OMe scale* (Octave-Median). The diameter of the circles corresponds to the duration of the syllables. The horizontal blue lines delimit the central octave surrounding the speaker’s median pitch.

ProZed [17], which is freely downloadable from the *Speech and Language Data Repository* at: http://sldr.org/sldr000778/en.

The diameters of the yellow circles correspond to the syllable durations and the dashed blue line corresponds to the Momel curve. The horizontal dotted lines correspond to the speaker’s median (middle line) and a half octave above and below the median, delimiting the speakers unemphatic pitch range corresponding to the median-centred octave. The values of the *Median* and the *Top* and *Bottom* of the central octave are given both in Hz and as musical notes as defined with respect to concert pitch at 440 Hz.

With this technique, the optimal parameters for the analysis of the fundamental frequency of the speaker are automatically determined from the median pitch.

6. Conclusions.

We propose in this paper that it is the octave, rather than the semitone, which should be considered the basic unit of a scale for natural speech prosody. This follows evidence reported in several studies based in neuronatomy, neurophysiology, behavioral studies, speech production as well as speech perception. In particular, we propose the use of the OMe scale to define an automatic display of the fundamental frequency curve. The reference (key) for such a scale is given by the median of the speakers fundamental frequency. The *Bottom* of the central octave of the speakers voice is consequently half an octave below the median while the *Top* is half an octave above.

The Bottom and Top lines of the display should not be thought of as physical obstacles for speakers. Obviously, in more spontaneous corpora we are likely to find a larger pitch range - up to two octaves has been reported in the literature. Pitch often goes beyond these lines, particularly in the case of the Top , but when it does so, it may be taken as a good sign that the speech is expressive or signalling important information.

Since the expressive use of pitch particularly concerns the top of the range, it is natural that the distribution of pitch in these cases will be skewed. It should be noted, however, that taking the median pitch rather than the mean pitch as reference value largely reduces the impact of the skewness, but this naturally remains to be tested on much more data.

Further research is also necessary on the variability of the median pitch for given speakers. As mentioned previously, the octave and half-octave intervals may be used in structuring the discourse (e.g. indicating focus, topic change, turn-taking) as well as in expressing arousal or attitudes. The relation between speakers’ span and median voice may have facilitated the perception of linguistic and paralinguistic functions of pitch range in speech production across genders, languages and speaking styles and may be interesting to examine in terms of prosodic universals. It opens up the debate on the formal and functional aspects of speech prosody as a result of learning and experience or of having some basis in the operation of the auditory system.

7. References

[1] Auran, C., Bouzon, C.; Hirst, D.J. “The Aix-MARSEC Project: An Evolutive Database of Spoken British English.” in Proceedings of the 2nd International Conference on Speech Prosody, Nara, Japan 2004.

[2] Bertrand, R.; Blache, P.; Espesser, R.; Ferre, G.; Meunier, C.; Priego-Valverde, B.; Rauzy, S. “Le CID -Corpus of Interactional Data-: protocoles, conventions, annotations.” in Travaux Interdisciplinaires du Laboratoire Parole et Langage d’Aix en Provence, 25, 25-55., 2007.

[3] Boersma, P.; Weenink, D. Praat, a system for doing phonetics by computer. http://www.praat.org [version 5.3.41, February 2013], 1992 (2013).

[4] Braun, M. “Speech mirrors norm-tones: Absolute pitch as a normal but precognitive trait.” Acoustics Research Letters Online 2, 8590.

[5] Braun, M. “A retrospective study of the spectral probability of spontaneous otoacoustic emissions: Rise of octave shifted second mode after infancy.” Hearing Research 215, 39-46. 2006.

[6] Braun, M.; Chaloupka, V. “Carbamazepine induced pitch shift and octave space representation.”, Hearing Research 210, 85-92. 2005.

[7] Carr, P.; Durand, J.; Pukli, M. “The PAC project: Principles and Methods.” Tribune des Langues Vivantes 36: 24-35. 2004.

[8] Chentir, A.; Guerti, M.; Hirst, D.J. “Extraction of standard Arabic micromelody. ” Journal of Computer Science, 5(2):8689, 2009.

[9] Delais-Roussarie, E.; Durand, J. Corpus et variation en phonologie du français: méthodes et analyses. Presses Universitaires du Mirail. 2003.

[10] De Looze, C. Analyse et interprétation de l’empan temporel des variations prosodiques en anglais et en français. Doctoral thesis, January 2009. LPL and Aix-Marseille University, Aix-en-Provence, France. 2009.

[11] De Looze, C.; Hirst, D.J. “L’echelle OME (Octave-MEdiane): une échelle naturelle pour la mélodie de la parole.” in Actes des XXVIIIes Journées d’Etude sur la Parole, Mons, Belgium, May 25-28 2010.

[12] Fant, G; Kruckenberg, A.; Gustafson, K.; Liljencrants, J.. “A new approach to intonation analysis and synthesis of Swedish.”, in Proceedings of the First International Conference on Speech Prosody, Aix en Provence. 283-286. 2002.

[13] Hirst, D.J. “Phonological implications of a production model of English intonation.”, Phonologica 1980, 195-201. 1981.

[14] Hirst, D.J. “Structures and categories in prosodic representations.”, in Cutler and Ladd (eds.) 1983. Prosody : Models and Measurements (Springer, Berlin), 93-109. 1983.

[15] Hirst, D.J. “A Praat Plugin for MOMEL and INTSINT with improved algorithms for modelling and coding intonation.”, in Proceeedings ICPhS 2007. 1233-1236. 2007.

[16] Hirst, D.J. “The analysis by synthesis of speech melody: from data to models.”, Journal of Speech Sciences, 1(1): 5583, 2011.

[17] Hirst, D.J. “ProZed: A speech prosody analysis-by-synthesis tool for linguists.”, in Proceedings of the 6th International Conference on Speech Prosody, Shanghai May, 15-18.

[18] Imig, T.J.;Morel, A. “ Tonotopic organization in the lateral part of posterior group of thalamic nuclei in the cat.” Journal of Neurophysiology 53: 836851 1985.

[19] ISO 16:1975. “Acoustics – Standard tuning frequency (Standard musical pitch).”, International Organization for Standardization. http://en.wikipedia.org/wiki/ISO_16 1975.

[20] Kuttner, F. A. “ Prince Chu Tsai-Yu’s life and work: a re-evaluation of his contribution to equal temperament theory.” Ethnomusicology, Vol. 19, No. 2 (May, 1975), pp. 163206.1975.

[21] Ladd, D.R. Intonational Phonology, (Cambridge University Press, Cambridge) 1996 [second edition 2008].

[22] Liu J.; Wang N.; Li J.; Shi B.; Wang H. “Frequency distribution of synchronized spontaneous otoacoustic emissions showing sex-dependent differences and asymmetry between ears in 2- to 4- day-old neonates.” International Journal of Pediatric Otorhinolaryngology. 2009 May; 73(5):731-6 2009.

[23] Morel, A. Codage des sons dans le corps genouillé median du chat: évaluation de l’organisation tonotopique de ses différents noyaux. PhD dissertation. Université de Lausanne. Juris, Zurich 1980.

[24] Morest, D.K. “The laminar structure of the medial geniculate body of the cat.” Journal of Anatomy 99: 143160 1965.

[25] Nolan, F. “Intonational equivalence: an experimental evaluation of pitch scales.” in Proceedings of ICPhS 15, Barcelona, 771-774. 2003.

[26] Paeschke, A; Sendlmeier, W. F. “Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements.” In Proceedings of the ISCA-Workshop ”On Speech and Emotion” (Belfast), SpeechEmotion-2000, 75-80. 2000.

[27] Wright, A. A. ; Rivera, J.J.; Hulse, S.H.; Shyan, M.; Neiworth,J.J. “Music perception and octave generalization in rhesus monkeys.”, Joural of Experimental Psychology Gen, Sep, Vol 129 No 3, 291-307 2000.

[28] Young, R. W. “Terminology for Logarithmic Frequency Units”, The Journal of the Acoustical Society of America. 11: 134. 1939.