Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Wavelet-based adaptation of pitch contour to Lombard speech

Juraj Šimko[1], *Antti Suni*[1,2], *Martti Vainio*[1]

[1]University of Helsinki, Finland, [2]Aalto University, Finland
{juraj.simko,antti.suni,martti.vainio}@helsinki.fi

Abstract

Increase in fundamental frequency (f_0) is one of the most robust and best-studied phenomena characterizing Lombard speech. In this work, three types of global transformation of f_0 contours from normal speech to Lombard condition are investigated: (1) a linear re-scaling of the quiet condition contour to match the mean and standard deviation of f_0 in Lombard speech, (2) a non-linear regression between the f_0 values in quiet condition against the corresponding f_0 values in the Lombard speech and (3) a multiple non-linear regression using components obtained by a wavelet decomposition of the quiet condition contours. The quality of fits is evaluated on a phonetically controlled corpus of Finnish sentences with varying prosodic focus and ambient noise conditions. The results show that the non-linear regression yields a smaller root mean squared error that the simple rescaling. Both methods are outperformed by the technique based on continuous wavelet transformation that uses hierarchical information encoded in speech signal. The findings are discussed in terms of their theoretical implications as well as their possible technological applications.

Index Terms: f_0 contour, Lombard speech, adaptation, continuous wavelet transform

1. Introduction

In a noisy environment, we speak louder, in general slower and with higher pitch than in a quiet place. While there is a large body of research of the relationship between f_0 level and variance in quiet and Lombard speech (speaking in an ambient noise) [1, 2, 3], relatively little is known about possible methods of adapting f_0 contours obtained for (or synthesized for) quiet speech to plausible realization of pitch in the same sentences in noisy conditions. The aim of this study is to partly fill in this gap by presenting and evaluating three possible methods of such transformation on prosodically rich speech data.

Linguistic and prosodic factors are known to influence f_0 changes in Lombard speech. Several studies have reported on differences between function and content words [4], and between stressed and un-stressed words [5] in f_0 adaptation to noisy background. A study using the same speech material as the present one has shown that linguistic signaling of focus is similar in Lombard speech to speaking in quiet environment although speakers modify their f_0 contours depending on the type of noise regardless of equal loudness [6]. The study also shows that the overall pitch movement during the utterances increases in response to the increase in overall noise level. Here we present a further analysis of f_0 modification in Lombard speech and also study the possible effects of prosodic and linguistic parameters such as focus, stress and sentence constituent.

One of the transformation methods presented here makes use of continuous wavelet transform (CWT) of f_0 contours. CWT transform captures inherent hierarchical nature of the analyzed signal. The analysis not only shows how information is distributed in time, but also reveals the possible interdependencies between the hierarchically organized speech constituents (see Fig. 1). Continuous wavelet transform is an invertible method; in essence it provides a decomposition of the analyzed signal to several components. Wavelet based decomposition of f_0 contours has recently been used to train a parametric statistical speech synthesis system [7].

2. Methods

2.1. Material

The corpus of phonetically controlled material used in this study consists of 11 Finnish declarative sentences recorded by 21 native Finnish subjects (11 females). In addition to a normal rendering of collected material in quiet surroundings, nine conditions with ambient noise (Lombard conditions) were collected: three types of noise (white, pink and babble) at three levels (60, 70 and 80 dB) were administered to subjects over headphones during recording sessions. The order of sentences and noise type/level was randomized. In the present study, only the quiet and 80 dB babble noise conditions are used.

Each sentence consists of three two-syllabic words: subject, verb and object, in that order. Every sentence has the same quantity pattern, with heavy first syllable of the subject, second syllable of the verb and both syllables of the object word. All syllables are of CV form with singleton consonants. The 11 sentences differ in segmental content and in prosodic form: 4 sentences were rendered with broad focus pattern, 3 with a narrow focus on the subject and the remaining 4 with a narrow focus on the object word. Each individual sentence was produced in both quiet and Lombard condition.

Fundamental frequency contours were extracted using praat [8], manually checked for outliers and corrected when necessary; f_0 measurements were excluded for portions with creaky voice. A gap filling procedure was used to fill in plausible f_0 contours spanning the unvoiced and excluded intervals (see [9] for details). Resulting f_0 values were converted from hertz scale to semitones in two steps: first, the mean f_0 in Hz from all broad focus sentences recorded in quiet condition was used as a base value for conversion, subsequently, the semitone values were further adjusted to achieve minimal possible differences among speakers in mean f_0 values in quiet and Lombard conditions.

In this work, only f_0 values from the vocalic intervals are used. More precisely, for every vocalic interval we extracted f_0 values from 10 equidistant time-points spanning the entire interval, using a cubic spline interpolation. This procedure yielded two sets of f_0 values, one from the quiet condition and another containing the values from the same sentences at corresponding time-points uttered by the same speakers in Lombard condition.

doi: 10.21437/SpeechProsody.2016-100

2.2. Three transformation methods

The following methods of transformation of the f_0 values quiet condition approximating the values in Lombard speech are used and evaluated in this paper.

2.2.1. Scaling

The first is a linear scaling of the quiet condition f_0 values to match the mean and standard deviation of pitch in Lombard speech: f_0 values for quiet condition were normalized to zero mean and unity variance (z-score) and subsequently multiplied by standard deviation of f_0 values in Lombard condition and increased by Lombard condition mean.

2.2.2. Third order regression

Second approach was to fit a third order linear regression model with the quiet condition values as an independent and the corresponding Lombard values as a dependent variable. The regression estimates were then used as an approximation f_0 in Lombard speech.

2.2.3. CWT-based regression

The last transformation is a third order regression of a CWT decomposition of f_0 contours from quiet speech to the (non-decomposed) corresponding contours from Lombard condition.

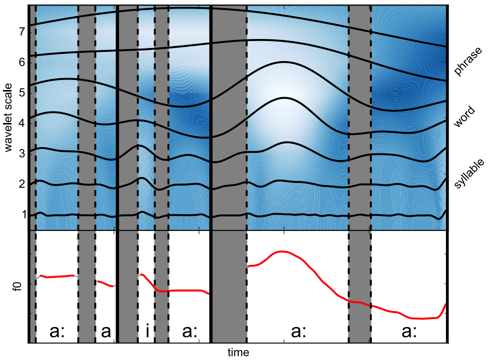

Figure 1: A CWT decomposition of an f_0 contour of a sentence with focus on the object word (in red) to seven components corresponding to wavelet scales separated by half an octave (in black). The heat-map represents continuous wavelet transform.

Continuous wavelet transform (Mexican hat mother wavelet) was used to decompose every quiet condition f_0 contour to seven components corresponding to wavelet scales separated by half an octave.

Fig. 1 presents an example of such decomposition and illustrates how wavelet analysis captures hierarchical nature of speech with components 1–3 primarily reflecting f_0 variation at syllable level, components 4–5 relating to word-level prosody (the latter especially reflecting the prominence of the first and the last word in the sentence) and components 6 and 7 capturing the phrase and sentence level phenomena like f_0 rise for the third (focused) word and overall pitch declination, respectively.

Ten equidistant f_0 values from each vocalic interval for every component of decomposition were obtained using cubic spline interpolation as in the original f_0 contours. Finally, a third order regression of the values from CWT components to corresponding Lombard speech values for fitted.

2.3. Factorization along prosodic/linguistic dimensions

It is plausible that the adaptation of f_0 contour to Lombard condition quantitatively or qualitatively depends on prosodic (stress, focus) and/or linguistic (subject, verb, object) characteristics of speech material. It is also possible that different speakers use different strategies to achieve this task.

To investigate these possibilities, the transformation techniques described above were in also applied on subsets defined according to prosodic and linguistic parameters: *focus*, lexical *stress* and *word* (subject, verb or object). The f_0 values were divided to those originating from the words under prosodic focus and those from unfocused words; to those from the first and the second syllable (Finnish has word initial lexical stress); and those from three different words. For each of these three divisions, the transformations were performed for each subset separately (two subsets for *focus* and *stress* factorizations and three for *word* division). Also, a factorization according to *all* three factors simultaneously was evaluated, leading to 10 subsets (the verb hasn't occurred under focus in our material).

The transformations were also performed for each speaker separately. Finally, factorizations along the prosodic / linguistic factors were carried out on data for individual speakers.

The quality of fits for each transformation is evaluated in terms of the root mean square error (RMSE) using the residuals of the fits, i.e., the differences between the estimates and the actual f_0 values in Lombard condition. For the factorized estimates, the residuals of individual fits obtained for individual prosodic / linguistic factors were combined for RMSE calculation. For the non-linear regression approaches (second and third types of transformation), combining residuals from partial fits is equivalent to obtaining residuals from a corresponding regression with the factors as interacting independent variables.

3. Results

Tab. 1 summarizes RMSE values for the three transformation methods and all prosodically and linguistically motivated factorizations, both with data for all speakers pooled together and separately. The quality of fits range from relatively poor with the error value as high as 3.21 for a simple overall fit using rescaling to very good for maximally factorized CWT-based fit (RMSE equal 0.72).

The results shows, that the CWT-based transformation provides better fits that the other two methods, with the non-linear regression being more precise than the simple re-scaling. Also, unsurprisingly, speaker dependent transformations lead to smaller errors than speaker independent ones. Taking prosodic and other factors into consideration naturally leads to more precise fits. Moreover, the lower RMSEs for factorization by *word* compared to that by *focus* and *stress* presumably reflects the fact that the data are divided to three subset for the former and only two groups for the latter factors.

More interestingly, the factorization has considerably greater influence on precision of speaker-divided CWT-regression transformation that on any other transformation method. For all other transformation types, including those using separate fits for individual speakers, the absolute improvement of fit (in terms of lowering RMSE) for *focus* and *stress* factorization ranges from 0.01 st for speaker independent rescaling to 0.1 st for speaker-divided regression. For *word* division, the improvement varies from 0.1 st (speaker-independent re-scaling) to 0.24 st (speaker-dependent regression) and for separate fits for *all* prosodic/linguistic factors from 0.2 to 0.6 st

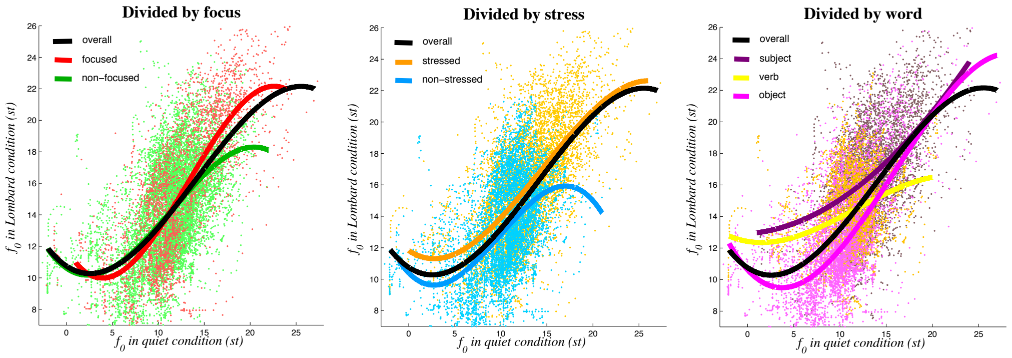

Figure 2: Third order fits (second transformation type) of Lombard f_0 values (y-axis) against corresponding quiet speech values (x-axis); data for all speakers are pooled together. Fits for all data points are shown in black, factorized regression fits for, from left to right, *focus*, *stress* and *word* are shown in different colors, see legends for each plot.

Table 1: RMSE values for three types of transformation and for each factorization; speakers pooled together (top) and separately (bottom). Δ and % Δ show by how much the factorization improves the estimates compared to the non-factorized transformation, as absolute difference in RMSE and as per cent of non-factorized error, respectively.

Factors:	**None**	**Focus**	**Stress**	**Word**	**All**
Scaling	3.21	3.20	3.18	3.11	3.01
Δ		0.01	0.03	0.10	0.20
% Δ		*0.2*	*1.0*	*3.1*	*6.3*
Regres.	2.82	2.79	2.74	2.60	2.31
Δ		0.03	0.08	0.22	0.50
% Δ		*1.1*	*2.8*	*7.8*	*17.8*
CWT	2.30	2.22	2.21	2.16	1.99
Δ		0.08	0.09	0.14	0.31
% Δ		*3.4*	*3.8*	*6.0*	*13.3*
Speaker +	**None**	**Focus**	**Stress**	**Word**	**All**
Scaling	2.31	2.25	2.23	2.14	1.86
Δ		0.06	0.07	0.17	0.44
% Δ		*2.5*	*3.1*	*7.4*	*19.2*
Regres.	1.97	1.88	1.87	1.74	1.37
Δ		0.09	0.10	0.23	0.60
% Δ		*4.5*	*5.1*	*11.7*	*30.2*
CWT	1.76	1.54	1.53	1.36	0.72
Δ		0.22	0.22	0.40	1.04
% Δ		*12.4*	*12.7*	*22.5*	*59.3*

(for the same two methods, respectively).

On the other hand, the improvements gained by factorization for speaker-dependent CWT-regression are 0.21, 0.22, 0.38 and 1.03 st for focus, stress, word and all division, respectively; that is almost twice as much as the next best improvement for the same factorization method.

The fact that the regression transformations provide better fits than the linear scaling methods suggests a non-linear influence of ambient noise on f_0 increase. Fig. 2 provides support for this insight and illustrates nature of this non-linearity. The figure shows the third order regression fits of quiet condition f_0s against the Lombard speech values – as used in the second transformation method – with factorizations along prosodic/linguistic factors (not divided by speakers). The non-factorized fit is shown in back, the factorized fits are plotted, from left to right, for *focus*, *stress* and *word* factorizations.

All fits show an inhibition of effect of ambient noise on f_0 transformation for low as well as high f_0 values (for the latter with an exception for the object word fit, the dark magenta curve in the last plot). This indicates that while there is an overall increase of f_0 in Lombard speech compared to quiet condition, the effect is attenuated at the lower (up to 5 st in our normalization) and higher (over 20 st) ends of f_0 range of quiet speech.

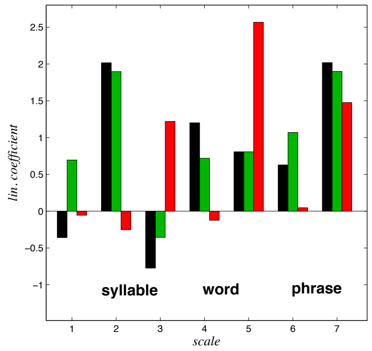

Figure 3: The values of linear coefficients of CWT-based transformation (speaker-independent factorization according to *focus* factor). Overall fit values in black, focused condition in red, unfocused in green.

All non-linearities suggested by the curves in Fig. 2 are statistically significant. More precisely, all 2nd and 3rd degree coefficients in the regressions used are significantly different from zero ($p < 0$), except the quadratic coefficient in factorization by *word* for subject words (violet curve in the third panel).

Turning our attention to the CWT-based fits, Fig. 3 shows the values of (seven) linear coefficients – that, presumably, best reflect the overall scaling in the regressions – in the cubic regressions of CWT-based decomposition of f_0 in quiet condition to f_0 contour of the same utterances in Lombard speech. The

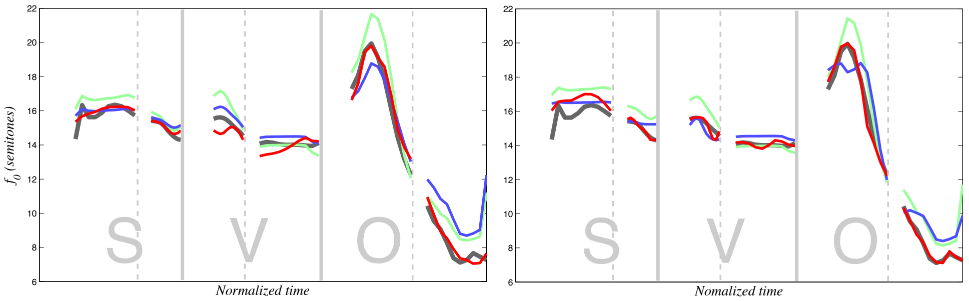

Figure 4: An example of f_0-contour transformation based on the investigated methods. Fits factorized according to all factors simultaneously, speaker independet (left) speaker dependent (right). Target Lombard contour in grey, re-scaling in green, non-linear regression in blue and CWT-based regression in red.

coefficients of the overall, non-factorized fit are shown in black, and the coefficients for the focused and non-focused subsets in red and green, respectively. The hierarchical speech components which best correspond to individual scales are also indicated in the figure (see also discussion related to Fig. 1 above).

As can be seen in the figure, the fits on non-focused and non-split material operate primarily on syllable and phrase level while the regression on focused material magnifies the decomposition component reflecting word level phenomena. In more detail, at syllable level, the regression on focused material acts more strongly at more slowly varied f_0 contours (scale 3 compared to scale 2) corresponding to longer syllables expected in focused context. At phrase level there is relatively small difference between the coefficients, the difference for scale 6 might be also related to overall tempo differences. The largest quantitative difference between coefficients is manifested for scales corresponding to words, that is precisely the hierarchical level at which focus is assumed to operate.

Finally, Fig. 4 provides an example of an adaption of quiet condition f_0 contour to Lombard speech using the investigated methods with factorization along *all* features simultaneously. The target Lombard contours are plotted in grey, transformation using re-scaling in green, non-linear regression in blue and CWT-based regression in red. In the left plot, the data are not divided by *speaker*, the right one shows complete factorization. As can be seen, the simple re-scaling exaggerates the (mostly upper) extrema; this is due to inability to scale the transformation to reflect the non-linearity of the relationship between normal and Lombard speech f_0. The CWT-based methods yield generally better fits that non-linear regressions – in particular in the focused part (object word in this example) – presumably due they sensitivity to hierarchical structure of speech.

4. Discussion

The evaluation show a progressively more precise transformation results from simple re-scaling to CWT-based regression. Mathematically, the RMSE cannot increase for more complex methods. The non-linear regression uses, in effect, a superset of independent variables compared to re-scaling (re-scaling is a form of linear regression). Similarly, as CWT yields a decomposition of an f_0 contour (a weighted sum of the components closely approximates the original contour), the quality of the fit using the components cannot be worse than that of the regression using original, not decomposed f_0 values.

Nevertheless, our subsequent analysis of the fits provides several theoretically interesting findings. The S-shape of the non-linear regressions shows a special nature of adaptation of f_0 contour to noisy conditions with "flattening" at both very high and very low intervals of f_0 values[1]. To our knowledge, this phenomenon has not been reported in the literature before. Although the general tendency is consistent with the "ceiling" effect discussed in [2] (f_0 increase as limited by physiological factors), our results show that the flattening effect is present in different prosodic context, even those (e.g., non-stressed syllables) where the speakers do not reach extremely high f_0 values. This finding suggest more complex efficiency influences at play where potential benefits of f_0 increase is traded-off against energy expenditure in a context dependent way.

A considerable improvement achieved using CWT-based hierarchical analysis and particularly the strong dependence of CWT fits on prosodic and linguistic context further emphasizes the complexity of pitch adaptation in adverse conditions. The way speakers seem to resolve communicative demands reflects prosodic hierarchies of speech: they are able to selectively magnify intonational phenomena associated with syllable, word or phrase level depending on their prosodic intentions (such as draw a focus to a particular word).

Of course, our relatively straightforward transformation methods cannot account for many other possible means of f_0-contour adaptation to noisy conditions such as temporal shifts of f_0 peaks and valleys within a syllable, word or entire utterance. We by no means claim that speakers adapt their intonation by a simple re-mapping of f_0 values used in normal speech to different, higher values when speaking in noise. We are currently exploring the effects of temporal adaptation (general slowing down) on transformation of pitch as well as of other acoustic dimensions such as energy and voice quality.

Despite these limitations, our methods, particularly the fully factorized CWT-based technique provides excellent precision and shows a potential to be used in technological applications such as speech synthesis. In order to evaluate its generalizability to different speakers and speech material, the evaluation methodology will need to be adapted to include more varied speech corpora and – unlike in this preliminary study – a separation of training and testing material. If successful, we will work on implementation of this adaptation method within an existing synthesis platform and use synthetic speech for perceptual evaluation of these and future adaptation methods.

[1] The non-linearity does not imply any particular relationship between the f_0 variance in quiet and Lombard conditions.

5. References

[1] J.-C. Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," *Speech Communication*, vol. 20, no. 1, pp. 13–22, 1996.

[2] ——, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *The Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.

[3] Y. Lu and M. Cooke, "The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication*, vol. 51, no. 12, pp. 1253–1262, 2009.

[4] R. Patel and K. W. Schell, "The influence of linguistic content on the lombard effect," *Journal of Speech, Language, and Hearing Research*, vol. 51, no. 1, pp. 209–220, 2008.

[5] C. L. Rivers and M. P. Rastatter, "The effects of multitalker and masker noise on fundamental frequency variability during spontaneous speech for children and adults." *Journal of Auditory Research*, 1985.

[6] M. Vainio, D. Aalto, A. Suni, A. Arnhold, T. Raitio, H. Seijo, J. Järvikivi, and P. Alku, "Effect of noise type and level on focus related fundamental frequency changes," in *Proceedings of Interspeech 2012*, 2012.

[7] A. S. Suni, D. Aalto, T. Raitio, P. Alku, and M. Vainio, "Wavelets for intonation modeling in HMM speech synthesis," in *8th ISCA Workshop on Speech Synthesis, Proceedings, Barcelona, August 31-September 2, 2013*, 2013.

[8] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2002.

[9] M. Vainio, A. Suni, and D. Aalto, "Emphasis, word prominence, and continuous wavelet transform in the control of HMM-based synthesis," in *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*. Springer, 2015, pp. 173–188.