Identifying Tenseness of Lombard Speech Using Phase Distortion

Gilles Degottex[1], *Elizabeth Godoy*[2] *and Yannis Stylianou*

University of Crete, Computer Science Dep. and FORTH, Inst. of Computer Science
Vasilika Vouton, 71110 Heraklion, Greece

Abstract

The "Lombard effect" describes speakers' tendency to increase their vocal effort when communicating in noise [1]. Most often, the Lombard effect is examined in terms of acoustic parameters such as pitch, duration, and spectral amplitude (e.g., tilt/slope, formants) [2, 3]. However, these parameters offer limited insight into voice quality, such as "tenseness" associated with increased vocal effort. Acoustically, one of the most significant indicators of tenseness relates to features of the glottal excitation signal: specifically, perceived tenseness of a voice is linked to a decrease in the glottal spectral tilt (i.e., slope) [4]. Unlike typical analyses of the Lombard effect, the work in [5] explicitly examines glottal source parameters. Unfortunately, glottal source estimation is a challenging and delicate problem. Consequently, the present work seeks to offer an alternative analysis framework that can also isolate contributions of the excitation source (from the vocal tract), but without explicit glottal source modelling. In particular, phase distortion (defined below) is used to highlight differences in voice quality, focusing specifically on the relative tenseness of Lombard speech compared to Normal speech.

The phase distortion is defined here as the group-delay with the influence of linear phase component removed. To calculate it, a harmonic model with pitch-synchronous analysis [6] is used. First, the minimum phase contribution of the vocal tract filter is removed from the instantaneous phase of the sinusoidal parameters. Then, using phase difference and antidifference operators, the linear phase component is also removed from the measurement. What remains after this processing is phase distortion, which depends only on the shape of the glottal signal. This computation has been already proposed for the estimation of glottal model parameters and emotional valence detection [7, 8]; that is, the link between this phase distortion and voice quality has been established. In the context of Lombard speech analysis, the variance of this phase distortion (calculated across 2 pitch periods and averaged with an ERB scale) is more interesting than the phase distortion itself, since it reveals the stability of the sinusoidal components in the time-frequency plan. Specifically, the boost of glottal source energy above aspiration noise levels at high frequencies (in voiced segments) can be clearly observed by examining the phase distortion variance (see Figure 1).

Figure 1 shows an example of aligned Normal and Lombard speech, along with the corresponding spectrograms and phase distortion variance. First, note that for the unvoiced,"noisy" parts of both the Normal and Lombard speech, the phase distortion variance is large for all frequencies. Now, considering the voiced speech segments, it can be seen that the phase distortion variance is large for the upper part of the spectrum (high frequencies) of the Normal speech and small for lower frequencies. On the other hand, for the Lombard speech, the phase distortion variance remains relatively low across all frequencies for the voiced segments. This effect can be seen most clearly comparing the segments before 0.4 and after 0.8 seconds. Referring back to previous discussions, the lower phase distortion variance indicates the boosting of glottal spectral energy above aspiration noise for tense voices. Thus, the phase distortion variance is effectively confirming the tenseness in the Lombard speech.

In addition to identifying tenseness in Lombard speech, the framework for the present analyses could potentially be extended to synthesize speech with different voice quality. That is, we observe the tenseness of Lombard speech in relation to decreased phase distortion variance. So, by reducing the phase distortion variance, can we synthesize a voice that is more tense (admittedly, a difficult task)? Furthermore, in terms of the Lombard effect, the present work clearly shows differences in voice quality between the Normal and Lombard speech, but these observations beg the question: how do differences in voice quality influence speech intelligibility?

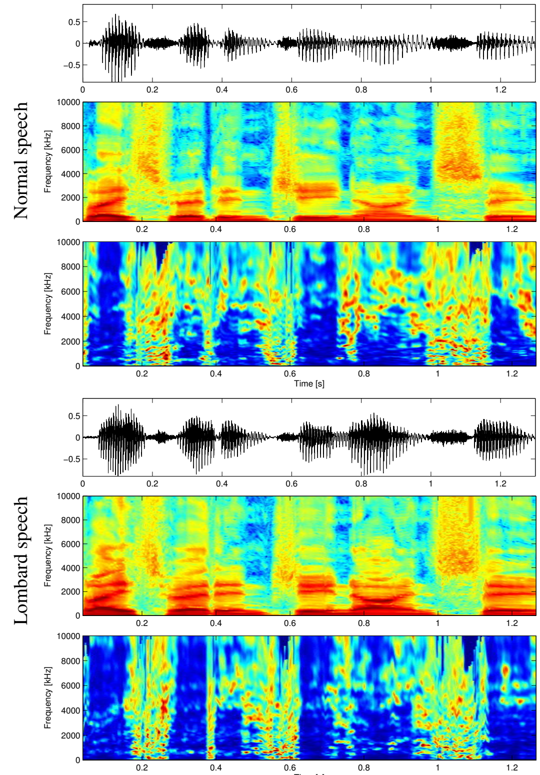


Figure 1: Lombard and normal speech samples with waveform, spectrogram and phase distortion variance

[1] Thanks to Swiss National Science Foundation (PBSKP2_134325).
[2] Thanks to LISTA project (E.U. FP7 FET-OPEN, 25623).

1. References

[1] Etienne Lombard, "The sign of the elevation of the voice," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, no. 2, pp. 101–119, 1911.

[2] J.C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *JASA*, vol. 93, no. 1, pp. 510–524, 1993.

[3] W. Van Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech producton: Acoustical and perceptual analyses," *J. Acous. Soc. Am.*, vol. 84, no. 3, pp. 917–928, 1988.

[4] G. Fant, "The LF-model revisited. transformations and frequency domain analysis.," *STL-QPSR*, vol. 36, no. 2-3, pp. 119–156, 1995.

[5] T. Drugman and Thierry Dutoit, "Glottal-based analysis of the lombard effect," in *Interspeech*, 2010, pp. 2610–2613.

[6] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AM-FM signal decomposition with application to speech analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 290–300, 2010.

[7] G. Degottex, A. Roebel, and X. Rodet, "Function of phase-distortion for glottal model estimation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 4608–4611.

[8] M. Tahon, G. Degottex, and L. Devillers, "Usual voice quality features and glottal features for emotional valence detection," in *Proc. International Conference on Speech Prosody*, 2012.

IDENTIFYING TENSENESS OF LOMBARD SPEECH USING PHASE DISTORTION

Gilles Degottex[1,2]*, Elizabeth Godoy*[2]*, Yannis Stylianou*[2]
degottex@csd.uoc.gr, egodoy@ics.forth.gr, styliano@ics.forth.gr
University of Crete, Computer Science Dep. and FORTH, Inst. of Computer Science

Thanks to
[1]Swiss National Science Foundation (PBSKP2_134325)
[2]LISTA project (E.U. FP7 FET-OPEN, 25623)

The Lombard Effect & Voice Quality

- The **Lombard effect** characterizes increased **vocal effort** of a speaker communicating in a noisy environment [1].
- The **tenseness** of Lombard speech is directly related to characteristics of the glottal excitation, specifically the **slope of the glottal spectrum**.
- Unfortunately, glottal source parameter estimation is a delicate and challenging problem [2].
- The **phase distortion** presents an alternative signal analysis (without explicit glottal source modeling) that **identifies voice quality**, such as tenseness.

Phase distortion and its variance

Signal Model

Full-band Adaptive Harmonic Model (aHM) (similar to [3])

$$s(t) = 2\Re\Big(\sum_{k=1}^{K} a_k(t) \cdot e^{jk\phi_0(t)}\Big) \qquad \phi_0(t) = \frac{2\pi}{f_s}\int_0^t f_0(\tau)d\tau$$

$a_k(t)$ complex function of amplitude and instantaneous phase interpolated from anchor values a_k^n estimated through LS solution at time t_n
We assume that a_k^n can be decomposed as in [4]:

$$a_k^n = e^{jkt_n} \cdot g_k^n \cdot c_k^n$$

e^{jkt_n} : Linear-phase dependent on the window position.
g_k^n : The glottal pulse shape
c_k^n : The Vocal Tract Filter (VTF)

Phase Distortion

Proposed for glottal model param estimation and emotional valence detection [4, 5] ⇒ **Link between Φ_k^n and voice quality** has been established.

$$\Phi_k^n = \angle\left(\frac{\tilde{a}_{k+1}^n / \tilde{a}_k^n}{\tilde{a}_1^n}\right)$$

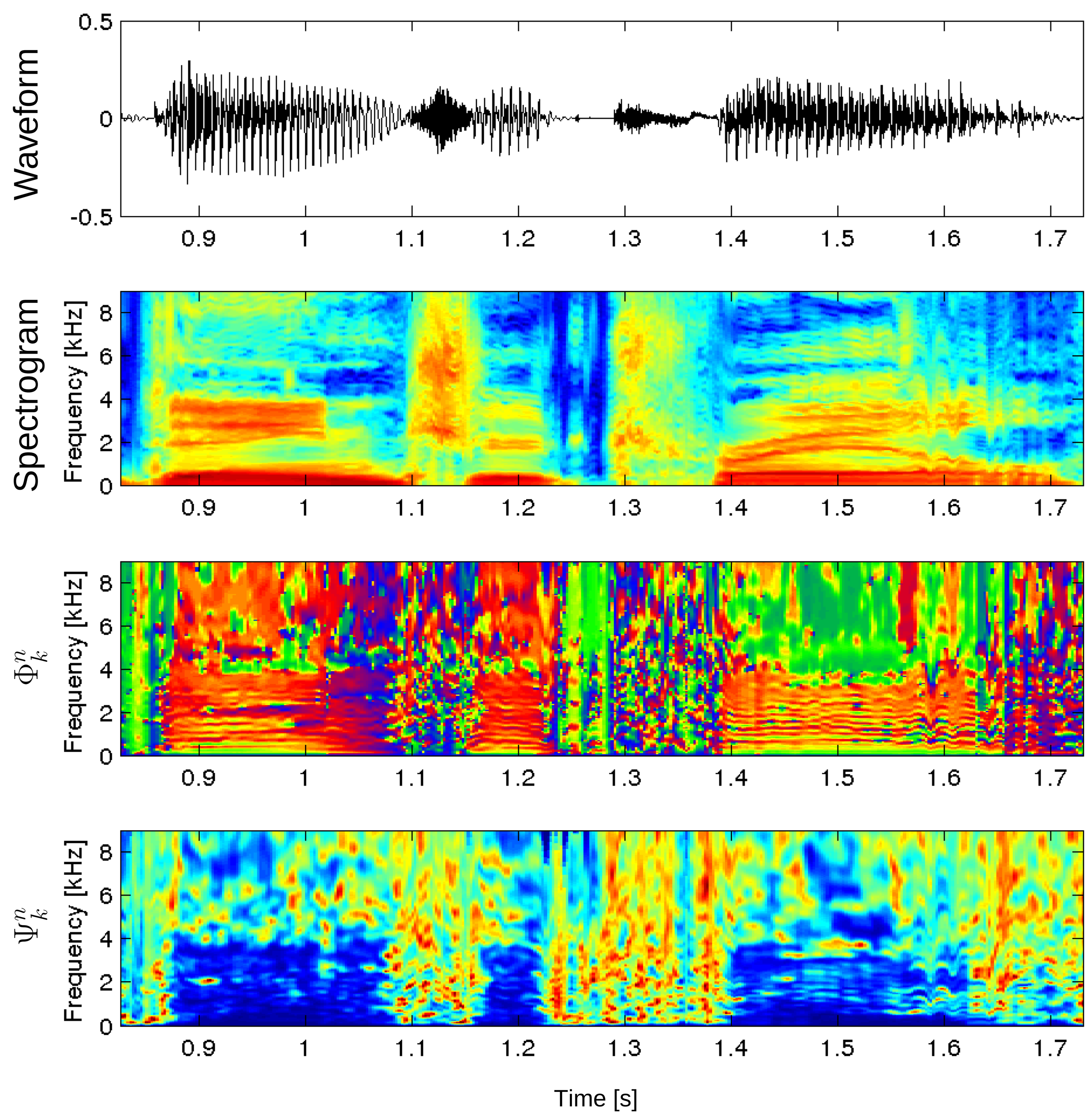
$\tilde{a}_k^n$ are the coefficients of the minimum-phase residual computed from a_k^n
1) $\tilde{a}_k^n$ is independent of the VTF amplitude
2) $\tilde{a}_{k+1}^n/\tilde{a}_k^n$ allows for a similar representation as the group-delay
3) The normalization by $\tilde{a}_1^n$ removes the influence of the window position t_n
⇒ Only the glottal pulse shape g_k^n influences Φ_k^n

Phase Variance

$$(\Psi_k^n)^2 = \text{var}_i(\Phi_k^n) = -2\log\left(\left|\frac{1}{M}\sum_{m\in B}\exp(j\Phi_k^m)\right|\right)$$

B set of Φ_k^n values in a 2 pitch period window around t_n ($M = |B|$).

Illustration of Phase Distortion



Low phase distortion **variance** ⇒ **deterministic** speech components.

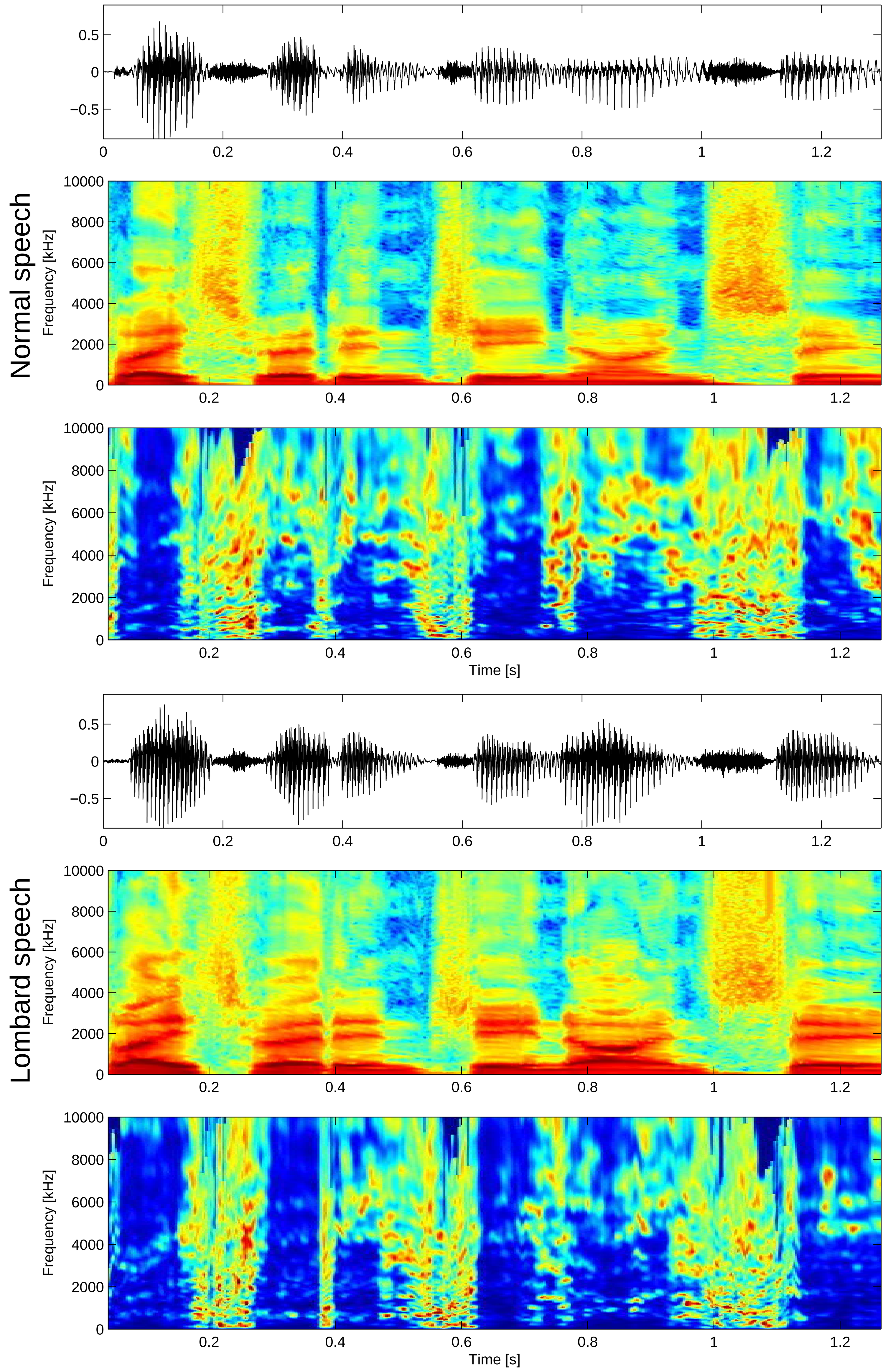
- Φ_k^n is relatively smooth in voiced segments at high frequencies compared to the erratic colors present in unvoiced segments ($t = 1.1$ and $t = 1.3$).

but still

- Ψ_k^n is low mainly below 4kHz in voiced segments.

Ψ_k^n: Normal vs. Lombard speech

- **Tenseness** of voice linked to the **flattening of the glottal spectrum** [6].
- **Lower Ψ_k^n** indicates boost of **glottal energy above aspiration noise**.
- Comparing [0.3s; 0.4s] and [0.8; 1] between Normal and Lombard speech, we therefore **observe** a **tenser** voice in the **Lombard** speech.



Perspectives

- To test on a large corpus.
- Could be extended to synthesize speech with different voice quality.
- How do differences in voice quality influence speech intelligibility?

[1] Etienne Lombard, "The sign of the elevation of the voice," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, no. 2, pp. 101–119, 1911.
[2] T. Drugman and Thierry Dutoit, "Glottal-based analysis of the lombard effect," in *Interspeech*, 2010, pp. 2610–2613.
[3] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AM-FM signal decomposition with application to speech analysis," *IEEE TASLP*, 2010.
[4] G. Degottex, A. Roebel, and X. Rodet, "Function of phase-distortion for glottal model estimation," in *ICASSP*, 2011.
[5] M. Tahon, G. Degottex, and L. Devillers, "Usual voice quality features and glottal features for emotional valence detection," in *ICSP*, 2012.
[6] G. Fant, "The LF-model revisited. transformations and frequency domain analysis.," *STL-QPSR*, vol. 36, no. 2-3, pp. 119–156, 1995.