Noise Reduction of Recorded Speech in an NMR Imaginer during Phonation for 3D Vocal Tract Modeling

[1]J. Přibil, [2]J. Horáček, [3]P. Horák, [1]I. Frollo

[1]Institute of Measurement Science, SAS, Dúbravská cesta 9, Bratislava, Slovakia
[2]Institute of Thermomechanics, AS CR, v.v.i., Dolejškova 5, Prague, Czech Republic
[3]Institute of Photonics and Electronics, AS CR, v.v.i., Prague, Czech Republic
Email: Jiri.Pribil@savba.sk

Abstract. *The paper presents two methods of noise reduction of recorded speech signal during phonation for the human vocal tract modeling in an NMR imaginer. The noise is mainly produced by gradient coils, have a mechanical character, and can be processed in spectral domain. Our first noise reduction method use real cepstrum limitation and clipping the "peaks" corresponding to the harmonic frequencies of mechanical noise. The second method is coming out from substation of the short-time spectra of two signals recorded withal: the first includes speech and noise, and the second is consisted of the noise only.*

Keywords: Speech processing, Noise reduction, NMR imaging

1. Introduction

Better knowledge of the inner structures of human body is enabled by using CT or NMR investigations. The noninvasive magnetic resonance scanning of vocal tract spaces of subjects for speech configuration or in phonation position of their resonant cavities for each vowel enables to develop the three-dimensional (3D) computer models of the vocal tract. The primary volume models of the human acoustic supraglottal spaces created from the NMR images can then be transformed into the 3D finite element (FE) models [1]. The FE models open new possibilities in simulating and understanding production of human voice and speech. The FE models of the human vocal tract enable to perform numerical simulations of voice production. These models should allow simulating pathological changes or voice quality variations due to slight geometry modifications of the human supraglottal acoustic space.

The purpose of future studies is to develop computational models of human vocal tract that allow more accurate representation of the 3D wave propagation, and especially real-time numerical simulations of phonation useful in modeling real clinical situations. The FE modelling enables also to simulate the influence of the acoustic impedance changes of the vocal tract by phonating into glass tubes or straws used in voice training and therapy in clinical practice [2]. The quality of the developed FE models has to be checked by a sufficiently accurate numerical simulation of the subject phonation during the NMR scanning and therefore the simultaneous acoustic recording of subject voice during the scan procedure is very important.

There exist several approaches to reduce the noise in speech. One group of these speech enhancement methods is based on the spectral subtraction of the estimated background noise [3]. The noise estimation techniques, usually coming out on statistical approaches [4], were not able to track the real variations in the noise thereby resulting in an artificial residual fluctuating noise and distorted speech. Other noise estimation techniques performed relatively better for stationary and slowly varying noise but showed degradations when the noise was non-stationary. For that reason we use another approach, based on cepstral speech modeling.

2. Reduction of NMR coil noise in speech signal using the cepstral model

In contradiction to other speech description principles (LPC etc.), the cepstral speech analysis is performed in the frequency domain. The cepstral speech synthesis (reconstruction of speech signal) is realized by a digital filter implementing approximate inverse cepstral transformation. For voiced speech the filter is excited by a combination of an impulse train and high-pass filtered random noise, for unvoiced speech the excitation is formed by a random noise generator. The transfer function of the vocal tract model is approximated by Padé approximation of the continued fraction expansion of the exponential function. The error of this inverse cepstral approximation depends on the number and the values of applied cepstral coefficients and the used approximation structure [5].

The fundamental frequency (F0) of voiced speech is represents by a typical peek in the real cepstrum, as well as the mechanical frequencies of gradient coils producing the noise into NMR imaginer. Our first noise reduction method is based on the limitation of the real cepstrum and clipping the “wrong peaks” corresponding to the harmonic frequencies of the mechanical noise. The second approach is based on the substation of the short-time spectra from two parallel processed signals: from the first microphone including the speech and noise, and from second microphone consisting only of the noise.

Single channel noise reduction by cepstrum clipping

Cepstral analysis of speech and noise signal is performed in the following way: from the input samples (after segmentation and weighting by a Hamming window) the complex spectrum by the FFT algorithm is calculated. In the next step the powered spectrum is computed and the natural logarithm is applied – see the block diagram in Fig. 1. Application of inverse FFT algorithm gives the symmetric real cepstrum. By limitation to the first N_0+1 coefficients, the Z-transform of the real cepstrum can be obtained. The truncated cepstrum represents an approximation of a log spectrum envelope

$$E(f)=c_0+2\sum_{n=1}^{N_0}c_n\cos(n\cdot 2\pi f) \qquad (1)$$

where the first cepstral coefficient c_0 corresponds to the signal energy.

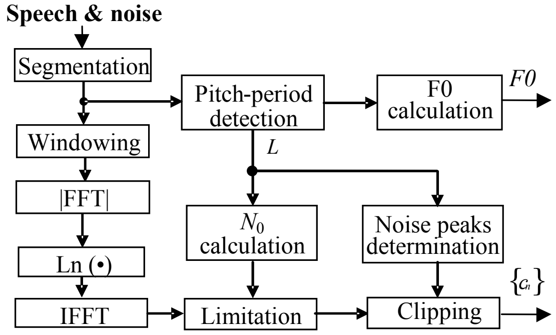

Fig. 1. Block diagram of cepstral analysis of the speech and noise signal.

The whole algorithm works in four steps:

1) Calculation of real cepstrum, pitch-period detection, F0 calculation.
2) Determination position of peaks in cepstrum corresponding to the frequencies of mechanical noise, and minimum number N_0 of cepstral coefficients (for sufficient log spectrum approximation [5]).
3) Limitation of real cepstrum and clipping peaks.
4) Reconstruction of input signal by the pitch-synchronous cepstral speech synthesizer.

Two channel noise reduction by spectral subtraction

Generally is the noisy speech signal *x(k)* interpreted as addition of a clean speech signal *s(k)* and a additive noise *n(k)*. The noisy signal is segmented and windowed to obtain a short time-frame of noisy speech. By transformation to the frequency domain using the FFT, we get the short-time spectrum $X(f, n)$. The enhanced speech spectrum can be obtained by subtracting a noise magnitude spectrum loaded by the microphone Mic. 2 from the noisy speech magnitude spectrum (of signal loaded by the first microphone Mic. 1:

$$S(f,n)=\left(\left|X(f,n)\right|-\left|N(f,n)\right|\right)\cdot e^{\varphi_n(f,n)}, \quad (2)$$

where $e^{\varphi_n(f,n)}$ represents the phase of noisy spectrum, and n is index of processed frame. On the resulting spectrum is applied natural logarithm and IFFT, whereby the limited real cepstrum is next obtained − see the block diagram in Fig. 2. The final clean signal is reconstructed also by cepstral speech synthesizer.

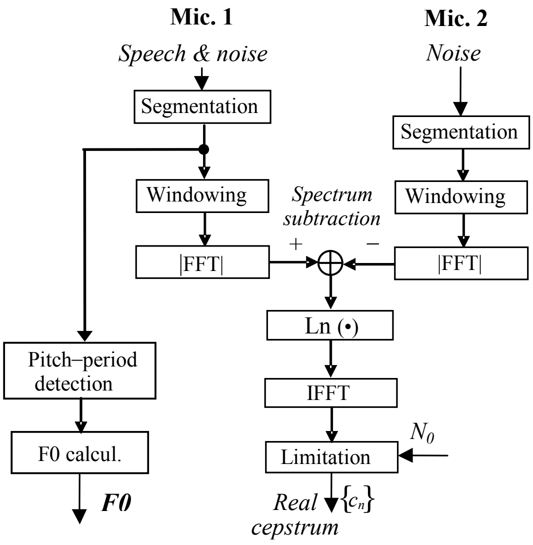

Fig. 2. Block diagram of two channel spectral subtraction method and cepstral analysis.

3. Experiments and Results

Our experiments were performed in the open-air 0.178T imaginer Esaote OPERA [6]. An arrangement of speech and noise recording measurement was following: the bed with testing person was set to 60 deg position (originated from the left corner – maximum rotation angle is 180 deg) − see Fig. 3. For the speech and RF coils noise signal recording, the front microphone (Mic. 1) was located on the 150 deg position; for recording the noise signal only, the back microphone (Mic. 2) was placed on the 30 deg position. Both microphones were located in the 10 cm above the bed (in the middle between both coils), and in the 60 cm radius of the central point of scanning area (RF coil). The measurement was realized with running scanning sequence: gradient echo, T1, coronal (TE = 18 ms), the background noise intensity (generated mainly by a temperature stabilizer) was I_0 = 55 dB (measured by DT-8820 device).

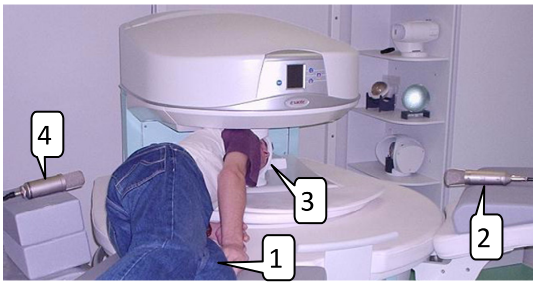

Fig. 3. An experiment arrangement: bed with testing person (1), front read microphone 1 (2), patient's head into RF coil (3), back read microphone 2 (4).

Speech as well as noise was recorded with the help of the M-Audio FireWire 1814 equipment connected to a personal computer through high-performance, high-resolution multi-channel interface of the IEEE 1394 (FireWire) bus. As the Mic. 1, the professional 1" Behringer dual diaphragm condenser microphone B-2 PRO (cardioids, omnidirectional or figure eight pickup pattern) was used. For Mic. 2 the RØDE NTK 1" condenser microphone with cardioid directional pattern was chosen. Signals from both microphones originally were recorded at 32 kHz, and resampled to 16 kHz. Collected database of speech and noise signals consists of 90 records of five separate phonated long vowels "a:", "e:", "i:", "o:", and "u:" from three male and three female non-professional speakers with mean time duration about 8 sec. The frame length depends on the mean pitch period L_0 of the processed signal. In our experiment, we use 24-ms frames for male voice, and 20-ms frames for female voice. The parameter for limitation of real cepstrum (chosen in correspondence on the period of noise part of signal

L_n), was set as $N_0 = 256$ (when $N_{FFT} = 1024$) for both voices; the minimum-phase cepstral coefficients $\{\hat{s}_n\}$ [5] were subsequently used for signal reconstruction.

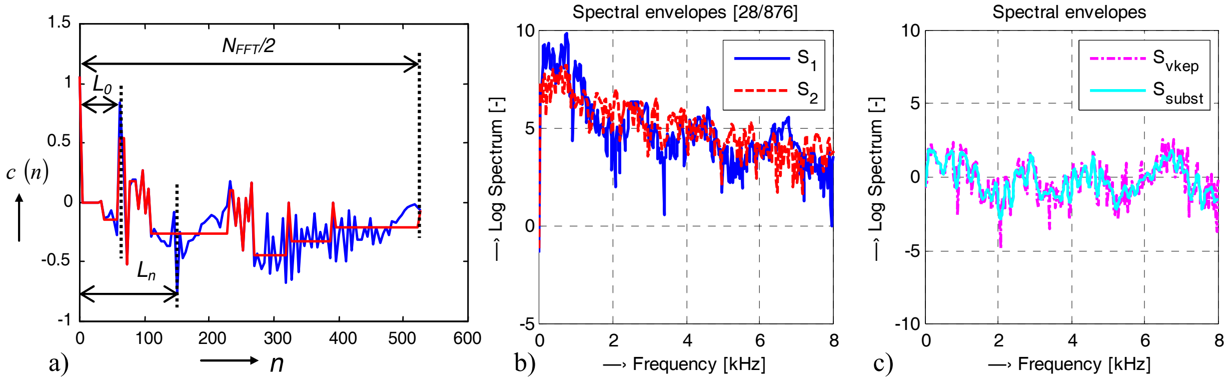

Fig. 4. Graphic examples of applied noise reduction method: real cepstrum clipping and limitation (a), two log spectral envelopes (b), resulting envelope after subtraction and cepstral reconstruction (c); processed speech and noise signal of vowel "a", female voice (F0 ≈ 220 Hz, processed 28th frame of total 876).

4. Conclusions

Performed experiment confirms usability of both applied noise reduction methods based on the cepstral speech model. The significant audible differences between the noisy and cleaned speech signal were observed for all of processed samples. Used professional condenser microphones have no effect on the homogeneity of low-magnetic field B_0 of the NMR imaginer, needed for obtaining the sufficient image quality. When the cardioid directional pattern is set for both microphones, the background noise can be ignored. From obtained results follows, that sufficient noise suppression can be reached by using of simple one microphone method (the two microphone approach is generally complicated for realization, and not bring any significant effect).

In the near future, we will use the listening tests for detail audio comparison of final speech signal cleared by these two methods. As an objective comparison criterion, the recognition score parameter of an Automatic Speech Recognition (ASR) system can be also applied too.

Acknowledgements

The work has been supported by the Grant Agency of the Slovak Academy of Sciences, project VEGA no. 2/0090/11, and by the Grant Agency of the Czech Republic (GA102/09/0989).

References

[1] Vampola, T., Horáček, J., and Švec, J.G., FE modeling of human vocal tract acoustic. Part I: Production of Czech vowels. Acta Acustica United Acustica, 94, 2008, 433-447.

[2] Laukkanen, A.M., Titze, I.R., Hoffman, H., and Finnegan, E.M., Effects of a semi-occluded vocal tract on laryngeal muscle activity and glottal adduction in a single female subject. Folia Phoniatrica et Logopaedica, 60, 2008, 298-311.

[3] Boll, S.F., Supression of Acoustic Noise in Speech using Spectral Subtraction. IEEE Trans. On ASSP, Vol. 27, No. 2, 1979, 113-120.

[4] Martin, R., Spectral Subtraction Based on Minimum Statistics. Proc. of EUSIPCO 1994, 1182-1185.

[5] Vích, R., Cepstral Speech Model, Padé Approximation, Excitation, and Gain Matching in Cepstral Speech Synthesis. Proc. of EURASIP Conference Biosignal 2000, 77–82.

[6] E-scan Opera. Image Quality and Sequences manual. 830023522 Rev. A, Esaote S.p.A., Genoa, April 2008.