INFORMATICA, 2006, Vol. 17, No. 2, 297–304

Discrimination of Homographs Distorted by a Lengthy Impulsive Noise [*]

Šarūnas PAULIKAS

Department of Telecommunications Engineering, Electronics Faculty
Vilnius Gediminas Technical University
Naugarduko 41-210, 03227 Vilnius, Lithuania
e-mail: sarunas.paulikas@el.vtu.lt

Dalius NAVAKAUSKAS

Department of Electronic Systems, Electronics Faculty
Vilnius Gediminas Technical University
Naugarduko 41-422, 03227 Vilnius, Lithuania
e-mail: dalius.navakauskas@el.vtu.lt

Received: March 2005

Abstract. The paper addresses the problem of discrimination of homographs when a lengthy segment of an uttered word is missing. The considered discrimination procedure is done by recognizer that operates on cepstrum coefficients extracted from the speech signal. For restoration of the missing speech segment rather than use of the known speech signal, it has been proposed to calculate speech signal characteristics: the period of fundamental frequency and intensity. By experimentation it has been shown that the polynomial approximation of speech signal characteristics improves homograph discrimination results. An extra computational burden associated with the proposed method is not high because it involves recalculation of the already extracted Fourier coefficients.

Key words: homograph discrimination, speech recognition, speech signal processing, speech signal characteristics, restoration, approximation.

1. Introduction

Speech recognition systems have reached the state of launching commercial products, however they still face a problem of maintaining a high recognition performance in adverse environments. The degradation of recognition performance is typically attributed to mismatch between training and testing conditions. Robust speech recognition methods include signal enhancement techniques as front-end and/or feature space transformations that reduce the variability due to noise (Gong, 1995). The inherent disadvantage of such techniques is that they make a poor assumption as to the nature of the noise or deal with a noise type difficult to model, which consists of short-time bursts (Vaseghi and Milner, 1997).

[*] For the support authors would like to thank prof. L. Ljung (Linköping University, Sweden).

Our framework deals with the long-time bursts of a random amplitude of impulsive noise or gaps in a speech signal that may extend to several hundred samples. Here instead of the replacement of cepstrum coefficients from the corrupted part of a speech signal by predicted ones (Potamitis *et al.*, 2001), we propose the method of reconstruction of the original cepstrum coefficients that uses the approximation of the intensity and fundamental frequency characteristics of a speech signal.

In Section 2, we present a formal derivation of the main equations needed to restore missing cepstrum coefficients of a speech signal. The results of experimentation on the recognition of isolated words from the Lithuanian language are given in Section 3. In the following, we consider only the approximation problem leaving aside determination of the position of a corrupted speech segment. However, during our case study we do use a modified version of the detector proposed in Vaseghi and Milner (1997).

2. Restoration of Missing Cepstrum Coefficients of a Speech Signal

The restoration of the distorted part of a voiced speech signal (see Fig. 1 for illustration) is based on the assumption that it is a quasi-periodical signal (Paulikas and Navakauskas, 1998; Paulikas and Navakauskas, 2003), that is why the distorted period is replaced by a modified version of the preceding undistorted period of a speech signal:

$$\hat{s}_{\mathrm{v}}(n) = \frac{I(m)}{I(m-1)} \cdot \hat{s}_{\mathrm{v}}\left(\left\lfloor \frac{T(m-1)\,(n - T(m-1))}{T(m)} \right\rfloor \right). \tag{1}$$

Here m is an index of the fundamental frequency period, $\lfloor\ \rfloor$ is an operation of rounding to minus infinity, $T(m)$ and $I(m)$ are the period of fundamental frequency and intensity characteristics, respectively (both constants considered here over one pitch period).

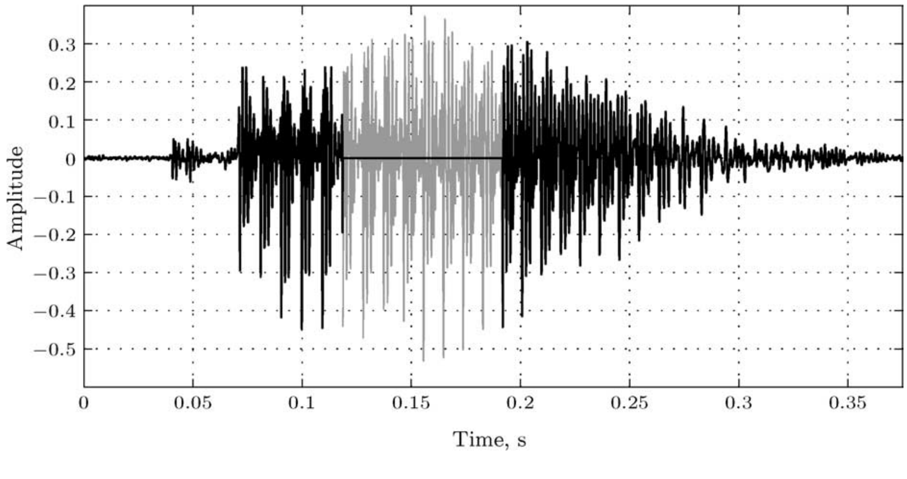
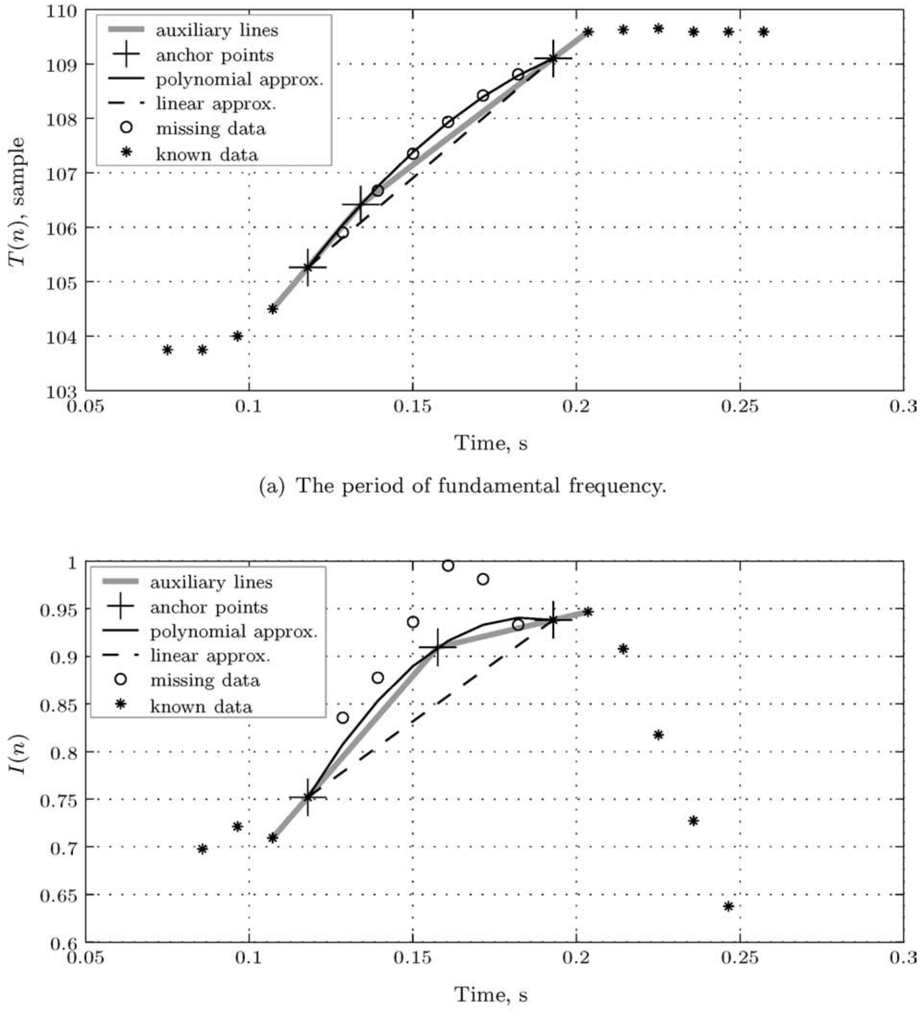


Fig. 1. Waveform of a distorted first syllable of the Lithuanian word "káltas" (black). The missing segment is superimposed (grey).

Computation of the necessary speech signal characteristics $T(m)$ and $I(m)$ (asterisks in Fig. 2) using an undistorted speech signal could be performed in numerous ways: for the solid exploration consult Hess (1983). Thus, it is more important to find the values of these characteristics for the missing segments (circles in Fig. 2). When a missing segment is short and speech characteristics do not vary a lot, it is plausible to approximate the missing points by line (dashed lines in Fig. 2). However, for the case of a lengthy missing segment, we do propose to use the second order approximation (solid line in Fig. 2). In order to compute it, at least three anchor points are required. Two of them can be given by



(a) The period of fundamental frequency.

(b) Normalised intensity.

Fig. 2. Characteristics of the first syllable of the Lithuanian word "káltas". Auxiliary lines are used to form anchor points, the latter ones – to form polynomial curve. The missing data are restored by linear and second order polynomial approximations.

the known values of characteristics on both sides of the distortion. The third anchor point is obtained by intersecting two auxiliary lines (grey lines in Fig. 2) that go through two known values of the characteristics taken from both sides of the distortion. Given three anchor points, we employ the conventional MSE criterion to fit the polynomial.

In the recognition process employed cepstrum coefficients of a speech signal are obtained by applying the inverse Fourier transform on the logarithm of the signal spectrum (Vaseghi, 2000; Higgins, 1990):

$$c(p) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln S(j\omega)\exp\left(j\omega p\right)\mathrm{d}\omega, \tag{2}$$

while the spectrum of the signal is calculated using the Fourier transform:

$$S(j\omega) = \int_{-\infty}^{\infty} s(n)\exp\left(-jwn\right)\mathrm{d}n. \tag{3}$$

In practice, cepstrum coefficients corresponding to particular period of fundamental frequency (m) can be obtained from a discrete cosine transform:

$$c^m(p) \approx \sum_{k=1}^{N} \ln S^m(k)\cos\left(\frac{p(k-1/2)\pi}{N}\right). \tag{4}$$

Here N is the total number of Fourier transform points, $S^m(k)$ is a magnitude of the short-time discrete Fourier transform of a segment of signal $s(n)$. Accordingly, cepstrum coefficients of restored part of the voiced speech signal (of m-th period of fundamental frequency) could be expressed as:

$$\hat{c}_{\mathrm{v}}^m(p) = \sum_{k=1}^{N} \ln \hat{S}_{\mathrm{v}}^m(k)\cos\left(\frac{p(k-1/2)\pi}{N}\right). \tag{5}$$

Here $\hat{S}_{\mathrm{v}}^m(k)$ is the restored magnitude of short-time discrete Fourier transform of the segment of voiced speech signal. It could be shown that $\hat{S}_{\mathrm{v}}^m(k)$ gets the following form

$$\begin{aligned}\hat{S}_{\mathrm{v}}^m(k) &= \frac{I(m)}{I(m-1)}\cdot\frac{T(m)}{T(m-1)}\cdot\hat{S}_{\mathrm{v}}^{m-1}\left(k\frac{T(m-1)}{T(m)}\right)\\ &\quad\times\exp\left(-k\frac{T^2(m-1)}{T(m)}\right),\end{aligned} \tag{6}$$

and corresponds to (1) expressed in frequency domain. Calculations in last expression are done for each short-time magnitude spectrum under restoration basing on previously found values, i.e., recursively. Magnitude spectrum of undistorted voiced speech signal serves as starting point of recursion.

The last two expressions formalize the proposed way of restoration of the missing cepstrum coefficients of a speech signal. Moreover, they enable us to control the restoration through the speech signal characteristics.

3. Experimentation on the Discrimination of Homographs

In order to verify the use of the proposed method, let us carry out recognition experiments on utterances of the Lithuanian word “káltas”.

The word is modelled using a hidden Markov model with 7 states, making use of the recognizer described in Cappé (2001). For the training phase of the model, we use the first 10 cepstrum coefficients obtained from 10 different utterances.

In the recognition experiments we use 3 additional (absent in the training set) utterances. The recognition is performed on artificially distorted utterances varying the place and duration of the distortion. To be more precise, the duration of the distorted segment is varied in length from 1 to 7 periods of fundamental frequency, while the beginning of the distorted segment varies approximately from 0.1 s to 0.2 s.

The distorted parts of utterances of the Lithuanian word “káltas” are reconstructed using (6), while the cepstrum coefficients are obtained from (5). For the reconstruction process, the necessary values of the fundamental frequency and intensity characteristics of the distorted segment are calculated, using linear as well as second order polynomial approximations.

The homograph of the Lithuanian word “káltas” is the word “kal̃tas”. Since these words differ only by the type of stress and possibly could be indistinguishable for the recognition system (especially in the cases of bad reconstruction), we also investigate the recognition of 3 additional undistorted utterances of the Lithuanian word “kal̃tas”.

In total 511 recognitions have been performed. Leaving aside the results that fall outside the acceptable recognition rate, the remaining results from 448 recognition experiments are summarized in Fig. 3.

Each point in the figure represents one word recognition experiment (indexed by i)

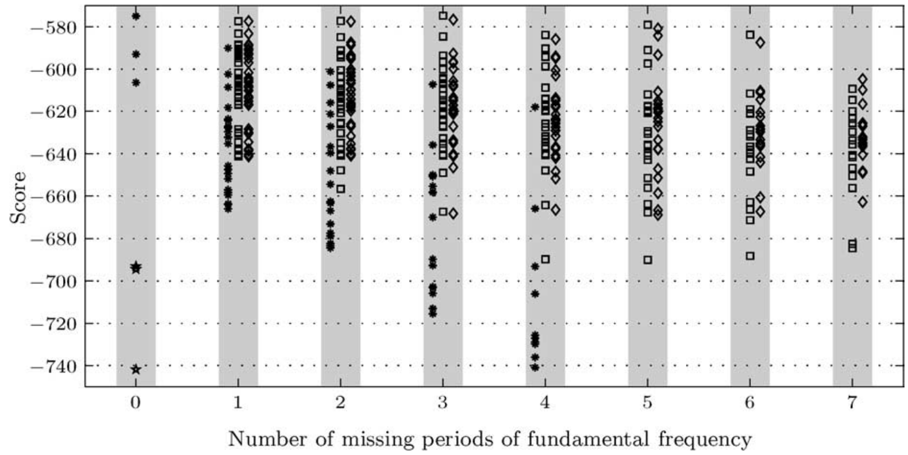


Fig. 3. Recognition results for the Lithuanian word “káltas”: without restoration (asterisks), employing linear approximation (squares), using 2nd order polynomial approximation (diamonds). For comparison (as pentagrams), the recognition results for the homograph – Lithuanian word “kal̃tas” are also shown.

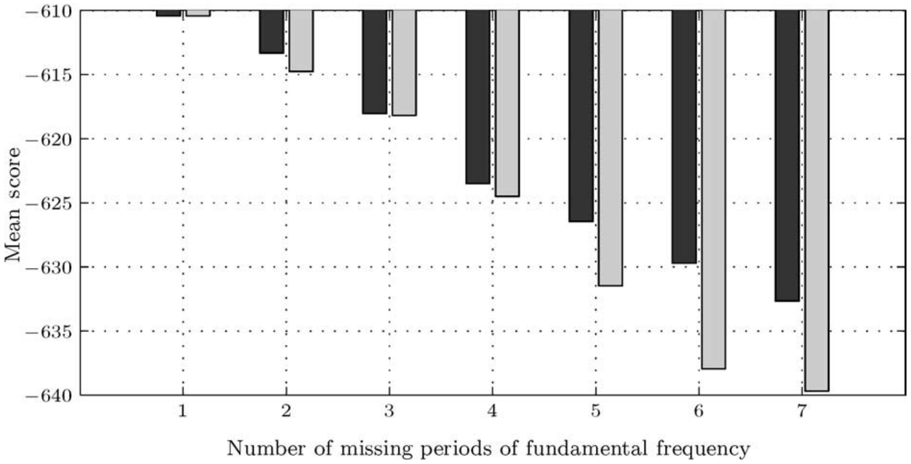


Fig. 4. Dependencies of averaged results of recognition of the Lithuanian word "káltas" on number of missing periods: employing linear approximation (light color bar), using 2nd order polynomial approximation (dark color bar).

that is evaluated by calculating the score value as follows:

$$S_i = \max_Q \left\{ \log P\left[Q, O_i|\lambda^{\text{káltas}}\right] \right\}, \tag{7}$$

here $Q = \{q_1, q_2, \ldots, q_t\}$ are all possible state sequences given the investigated observation O_i, using model $\lambda^{\text{káltas}}$.

As expected, the recognition results vary not only for different utterances – they also depend on the place and duration of a missing segment. It can be seen that when the length of the missing segment increases, the recognition results for the utterances of the word "káltas" without restoration soon reach similar score values of the word "kal̃tas". The behaviour that is more complicated and less dependent on the length of missing segment is seen from the recognition experimentation data for the restored signals. Experimentation data show increase in discrimination of considered homographs and confirm usefulness of the proposed restoration procedure.

In order to reveal the differences, let us examine the average values of the corresponding scores (see Fig. 4). The results in the range from 1 to 4 periods of fundamental frequency are similar when using both approximations, however, starting from the 5th period, employment of only the second order approximation starts to improve the recognition performance.

4. Conclusions

The presented restoration method of missing cepstrum coefficients increases the discrimination of homographs in the cases where the speech signal is corrupted by a long-term impulsive noise or long duration gaps are present.

It has been shown by the experimental study, that the approximation of fundamental frequency and intensity characteristics, using only the second-order polynomial as compared with the linear approximation, improves the recognition results when the duration of a missing segment is longer than 5 periods of the fundamental frequency.

The method considered deals with the frames of Fourier coefficients already extracted for the recognition purpose, i.e., to compute cepstrum vectors. That is why the recalculation of Fourier coefficients gives only little computation overhead to the whole speech recognition process. An extra time, naturally, is spent to compute and approximate the intensity and pitch characteristics. Thus, only simple and low order approximations were investigated.

Acknowledgements

Research in part was supported by The Swedish Institute New Visby project Ref. No. 2473/202(381/T81).

References

Cappé, O. (2001). H2M: A set of MATLAB/OCTAVE functions for the EM estimation of mixtures and hidden Markov models. *H2M Toolbox Manual*, Ver. 2.0, p. 16.

Gong, Y. (1995). Speech recognition in noisy environments: a survey. *Speech Communication*, **16**, 261–291.

Hess, W. (1983). *Pitch Determination of Speech Signals*. Springer-Verlag Berlin and Heidelberg GmbH & Co. KG.

Higgins, R.J. (1990). *Digital Signal Processing in VLSI*. Prentice-Hall, Inc.

Paulikas, Š., and D. Navakauskas (1998). Restoration of localized pitch and intensity variations of speech signals. In *Proceedings of the 1st International Conference Digital Signal Processing and its Applications DSPA'98*, Moscow, Russia, vol. 1. pp. 130–136.

Paulikas, Š., and D. Navakauskas (2003). New implementation scheme for the restoration of voiced speech signals. *Informatica*, **14**(3), 349–356.

Potamitis, I., N. Fakotakis and G. Kokkinakis (2001). Robust automatic speech recognition in the presence of impulsive noise. *Electronics Letters*, **37**(12), 779–780.

Vaseghi, S.V. (2000). *Advanced Signal Processing and Digital Noise Reduction*. John Wiley & Sons Ldt. and B. G. Teubner, Great Britain, 2 edition.

Vaseghi, S.V., and P. Milner (1997). Noise compensation methods for hidden Markov model speech recognition in adverse environments. *IEEE Trans. on Speech and Audio Processing*, **1**(5), 11–21.

Š. Paulikas was born 1969 in Vilnius, Lithuania. Received radioelectronics engineer diploma with honor in 1992, MSc in electronics degree in 1994 and doctor of electrical and electronical engineering degree in 1999. Presently working as associate professor at Telecommunications Engineering Department of Electronics Faculty at Vilnius Gediminas Technical University. Main research interests are digital communications and speech signal processing.

D. Navakauskas is professor at Electronics Systems Department of Vilnius Gediminas Technical University. He received honor diploma of radioelectronics engineer in 1992, MSc in electronics degree in 1994, doctor of electrical and electronical engineering degree in 1999, passed habilitation procedure in 2005, all at Vilnius Gediminas Technical University. His main research interests include artificial neural networks, speech signal processing and nonlinear signal modeling.

Homografų, sugadintų ilgu impulsiniu triukšmu, skyra

Šarūnas PAULIKAS, Dalius NAVAKAUSKAS

Straipsnyje nagrinėjama homografų skyros problema, kai ilga ištarto žodžio dalis yra sugadinta ar visai prarasta. Homografo identifikavimą atlieka automatinė kalbos atpažinimo sistema, kurios veikimas pagrįstas kalbos signalo kepstro koeficientais. Parodoma, kad kepstro koeficientų iš prarasto žodžio segmento atstatymas taikant kvadratinį kalbos signalo pagrindinio tono ir intensyvumo aproksimavimą pagerina žodžio atpažinimą lyginant su įprastiniu tiesiniu šių charakteristikų aproksimavimu. Papildomų skaičiavimų sietinų su siūlomu metodu, apimtis nėra didelė, kadangi juose modifikuojami atpažinimo sistemos rasti Furjė koeficientai.