Speech Features Extraction Using Cone–shaped Kernel Distribution

Janez Žibert, France Mihelič and Nikola Pavešić

Faculty of Electrical Engineering, University of Ljubljana,
Tržaška 25, SI–1000 Ljubljana, Slovenia
{janez.zibert, mihelicf, nikolap}@fe.uni-lj.si
http://luks.fe.uni-lj.si

Abstract. The paper reviews two basic time–frequency distributions, spectrogram and cone–shaped kernel distribution. We study, analyze and compare properties and performance of these quadratic representations on speech signals. Cone–shaped kernel distribution was successfully applied to speech features extraction due to several useful properties in time–frequency analysis of speech signals.

1 Introduction

Joint time–frequency signal representations characterize signals over the time–frequency plane. They combine time domain and frequency domain analyzes to yield a potentially more revealing picture of the temporal localization of signal's spectral components.

Due to varying degree of nonstationarity ranging from plosive bursts to vowel voicing speech is a complex signal and it is a tough challenge to obtain a time–frequency representation with a satisfying time and frequency resolution.

We are proposing a new modified method of speech features extracting based on mel–frequency cepstral coefficients with use of the cone–shaped kernel distribution. We are additionally studying dynamic features, which are modelled from basic parameters obtained by the new method. We are investigating several estimates of the time derivatives approximated by regression coefficients and coefficients determined by trigonometric functions.

Analyzes and tests are performed for different sets of speech features obtained from spectrogram and cone–shaped kernel distribution using speech recognition system based on hidden Markov acoustic models (HMM).

Our main goal has been to incorporate different time–frequency distributions into a speech features extraction process and potentially find an alternative way of deriving speech features based on these distributions.

2 Cone–shaped Kernel Distribution

The cone-shaped kernel distribution (Zhao–Atlas–Marks distribution) [2] is a member of Cohen's class of time–frequency distributions [1]. The spectrogram

which is almost exclusively used for the analysis of speech signals, is also a member of this class of the quadratic representations.

A time–frequency distribution (TFD) of Cohen's class is given by

$$C_x(t,f;\Pi) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \Pi(s-t,\xi-f)W_x(s,\xi)ds\,d\xi, \tag{1}$$

where $W_x(s,\xi)$ is a Wigener–Ville distribution [3] of the signal $x(t)$ [1], [4]. The alternative definition of Cohen's class distributions can be interpreted as the two dimensional Fourier transform of the ambiguity function [3] $A_x(\xi,\tau)$ multiplied by the kernel $k(\xi,\tau)$:

$$C_x(t,f;k) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} k(\xi,\tau)A_x(\xi,\tau)e^{-j2\pi(f\tau+\xi t)}d\xi d\tau. \tag{2}$$

The kernel function $k(\xi,\tau)$ defines characteristics of the given distribution and provide a good way to design a representation with desired properties.

For kernel $k(\xi,\tau) = A_h^*(\xi,\tau)$, where $A_h^*(\xi,\tau)$ is a complex conjugate of ambiguity function of analysis window h, we obtain a spectrogram, which can be interpreted as a quadrat of a short–time Fourier transform:

$$S_x(t,f) = \left|\int_{-\infty}^{\infty} x(u)h^*(u-t)e^{-j2\pi fu}du\right|^2. \tag{3}$$

If we impose to the distribution from the Cohen's class the condition to preserve time– and frequency–supports [4], [5], one of the choices is kernel function $k(\xi,\tau) = \frac{\sin(\pi\xi\tau)}{\pi\xi\tau}$, which defines Born–Jordan distribution [4]. If we further smooth this distribution along the frequency axis, we obtain Zhao–Atlas–Marks distribution [2], defined as

$$CKD_x(t,f) = \int_{-\infty}^{+\infty} h(\tau)\cdot\left[\int_{t-|\tau|/2}^{t+|\tau|/2} x(s+\tau/2)x^*(s-\tau/2)ds\right]e^{-j2\pi f\tau}d\tau. \tag{4}$$

Typically the smoothing window can be expressed as

$$h(\tau) = \frac{1}{\tau}\exp(-\alpha\tau^2), \tag{5}$$

which determines a cone–shaped kernel function in the ambiguity plane, hence the name the cone–shaped kernel distribution (CKD).

2.1 Comparison with Spectrogram

Spectrogram (3) and CKD (4) are members of Cohen's class of energy distributions [1]. They preserve energy of a signal over the time–frequency plane and are covariant by translations in time and in frequency [1]. All quadratic TFDs satisfy quadratic superposition principle [3] from where we get so called cross–terms which can be identified as a phenomena of interference in the time–frequency plane.

The interference terms of the spectrogram are restricted to those regions of the time–frequency plane, where corresponding signal terms overlap. Hence, if two signal components are sufficiently far apart in the time–frequency plane, then their cross–terms will nearly be identical zero. This property, which is a practical advantage of the spectrogram, is in fact a consequence of the spectrogram's poor resolution, [1], [3], [5]. The resolution depends of the given analysis window. If we used shorter (longer) window h in (3), we would obtain better (poorer) time and poorer (better) frequency resolution [5]. This is direct consequence of the uncertainty principle [6].

In general, it could be shown that there exists a general trade–off between good time–frequency resolution and quantity of interference terms [3]. The Wigener–Ville distribution (WVD), on the other hand, has excellent time–frequency concentration owing to number of good mathematical properties, but it also possesses substantial interference terms. The main idea in of deriving other distributions from Cohen's class is to smooth WVD (ambiguity function) with a kernel (2) by preserving time–frequency resolution and reducing interference terms. If we choose for the kernel $k(\xi, \tau)$ decreasing function of product $\xi \cdot \tau$, we obtain low pass function in ambiguity plane, which reduces interference terms. Cone–shaped kernel distribution is designed from the kernel $k(\xi, \tau) = \frac{\sin(\pi\xi\tau)}{\pi\xi\tau}$, which has above mentioned characteristics.

2.2 Time–frequency Representations of Speech Signals

Fig. 1 compares a spectrogram and cone-shaped kernel distribution for a short speech segment consisting of two pitch periods of a voiced Slovene vowel `/e/` spoken by a male speaker. The speech signal was sampled at a rate of 16 kHz and speech waveform represents approximately 22 ms of the sound.

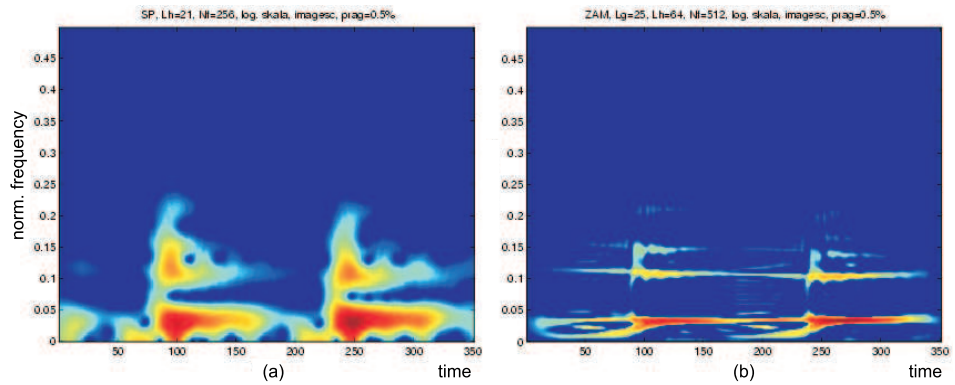

Fig. 1. Comparison of the spectrogram (a) and the cone–shaped kernel distribution (b) of a speech signal of the Slovene vowel /e/. The figures are produced using Time–frequency Toolbox [8].

Spectrogram in Fig. 1(a) is everywhere nonnegative. The representation of the analyzed signal is relatively clear thanks to its excellent cross–term prop-

erties. The main disadvantage of the spectrogram is that simultaneous good resolution in time and frequency is not possible. In spectrogram in Fig. 1(a) we chose shorter analysis window, so we obtained better time and poorer frequency resolution. Formant frequencies are smeared although they can be localized at every pitch period during the vowel voicing (approximately every 10 ms). There is another disadvantage of spectrogram, which can not be seen from Fig. 1(a). Since spectrogram does not satisfy the marginal and the finite time support property [4], spectral content in the representation of instantaneous changes in speech signals (e.g. bursts) is smeared in time and spanning over the entire spectral domain [7].

In Fig. 1(b) a representation by the CKD is shown, where the kernel function (5) with $\alpha = 1$ was used. CKD shows located formant frequencies much more precisely than the spectrogram. Note that spectral energy of formant tracks are smeared in the time direction, which is a direct consequence of using additional smoothing window (filter) in CKD to eliminate interference terms. There is another disadvantage of most of the Cohen's class distributions: they do not satisfy fundamental property of nonnegativity due to the interference terms. As a direct consequence we had to introduce additional threshold levels to suppress negative values in order to improve the representation of the analyzed signal [5].

3 Speech Features Extraction

In previous section the spectrogram and the cone–shaped kernel distribution was compared for needs of speech signal analyzes. We have shown differences and revealed some good characteristics of the CKD, which had motivated us to apply this time–frequency distribution to speech features extraction using mel–frequency cepstral coefficients.

3.1 Mel–Frequency Cepstral Coefficients

Our approach for deriving mel–frequency cepstral coefficients using CKD is similar to a basic method using spectrogram as a front–end.

Firstly some simple preprocessing operations to speech signals were applied prior to performing actual signal analysis: DC mean removal from the source waveform and pre–emphasis filtering due to physiological characteristics of the speech production system [9].

Then a time–frequency transformation of speech signal was performed. Given a discrete–time signal $s[n]$, $n = 1, \ldots, N$, we obtained time–frequency representation of dimension $N \times K$, where K is a number of frequency points. The representation can be rewritten in a matrix form $D \in \mathbf{R}^{N \times K}$. In case of the spectrogram D was computed from (3), hence $D = S[n,k]_{n=1,k=1}^{N,K}$, and in case of the CKD $D = CKD[n,k]_{n=1,k=1}^{N,K}$ was calculated from (4).

Next, a mel–filter bank analysis was performed. The bank of triangular filters equally spaced along the mel-scale frequency resolution [9] was introduced in a

matrix form. If M is the matrix of filters, the filtering could be implemented as a product of matrices M and D followed by a logarithmic operation:

$$M_{TF} = \log(M \cdot D), \tag{6}$$

where $log(\cdot)$ means element–by–element log operation. Elements $m[n, q]$, $q = 1, \ldots, Q$ (Q is a number of filters) of matrix M_{TF} present log filterbank amplitudes of signal $s[N]$ in time n modelled with different time–frequency distributions.

In order to derive cepstral coefficients the discrete cosine transform (DCT) was used

$$c[i, n] = \sqrt{\frac{2}{Q}} \sum_{q=1}^{Q} m[n, q] cos\left(\frac{\pi i}{Q}(q - 0.5)\right), \tag{7}$$

where $Q = 32$. In each frame of 10 ms length 12 cepstral coefficients were obtained appending the logarithm of the signal energy.

3.2 Modelling Dynamic Features

It is widely known that the performance of speech recognition system can be greatly enhanced by adding time derivatives to the basic static parameters [10].

According to the procedure described in previous subsection there were used different modelling approaches to estimate dynamic features from time–frequency representation of speech signals.

First approach was to perform regression analysis, where first and second order time derivatives were approximated by regression coefficients. The feature set consists from the basic cepstral parameters computed from the CKD with added the first and the second order regression coefficients is called CKD1. This is a standard approach. Derived features were used as a reference set in our further experiments.

Second approach of deriving dynamic features was estimation of coefficients of function arctan. This was performed by fitting each set i of basic features with function

$$f_{i,n}(x) = a_1^{(i,n)} \arctan(x) + a_0^{(i,n)} \tag{8}$$

for $n \in [-N_d, N_d]$, where N_d is a number of feature vectors. We achieved this by minimizing

$$\sum_{m=-N_d}^{N_d} |c[i, m] - f_{i,n}(x_m)|^2, \tag{9}$$

where x_m, $m = -N_d, -N_d + 1, \ldots, N_d$ were symmetrically but not equally chosen from interval $[-2.5, 2.5]$. The coefficients $a_1^{(i,n)}$ were added to basic mel–frequency cepstral coefficients. This kind of features was signed as CKD2.

A similar approach was used to produce features CKD3. Here dynamic features were derived by fitting each set of basic data with function

$$g_{i,n}(x) = a_0^{(i,n)} + a_1^{(i,n)} \cos(x) + a_2^{(i,n)} \sin(x). \tag{10}$$

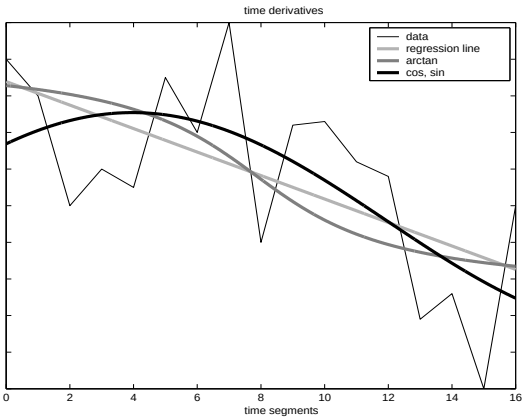

Fig. 2. Different approaches in modelling dynamic features from cone–shaped kernel distribution.

This is a slightly different approach as one used in [11]. The coefficients $a_1^{(i,n)}$ and $a_2^{(i,n)}$ obtained by minimizing (9) with function $g_{i,n}(x)$, where x_m were chosen equally from interval $[0, \pi]$, were appended to basic parameters to form third set of features.

Fig. 2 shows different approaches in modelling dynamic features from basic parameters based on the CKD.

4 Speech Recognition

4.1 Phone Recognizer

A Slovene speech database K211d [12] was used for evaluation purposes. Speech database K211d is a multi-speaker isolated-word corpus designed for phonetic research studies of the Slovene spoken language. The K211d lexicon consists of 251 carefully selected words to provide a representative sample of all Slovene allophones. Ten speakers (5 female + 5 male) uttered all 251 words stored. The corpus consists of 16,947 phones derived from 32 different allophones [12]. We used speech data from 6 speakers (3 female + 3 male) for training purposes. Test part (the rest of the data) includes 8848 phones.

Phone recogniser was based on HMMs. The acoustic models were standard left–to–right HMMs consisting of three states and three output Gaussian probability functions modelled with diagonal covariance matrices. Context–independent phonetic HMMs were trained using the HTK toolkit [13].

4.2 Phone Recognition

Phone recognition was performed using identical conditions. This means that simple HMM topology for context–independent phone recognition with equal number of parameters was used through all experiments testing different kind of features.

Features named SPEC were selected as the reference set of features. The SPEC feature set included mel–frequency cepstral coefficients based on spectrogram adding first and second derivatives derived from regression analysis.

Table 1. Phone recognition results with different sets of speech features.

	SPEC	CKD1	CKD2	CKD3
accuracy	78.25%	77.97%	70.18%	78.19%

The phone recognition results for different feature sets are shown in Table 1. The phone recognition results of SPEC, CKD1 and CKD3 are almost identical. This was expected for the SPEC and CKD1 features based on similar modelling techniques only with different time–frequency distributions. Poorer results were obtained with CKD2 features. This kind of features include just estimations of the first order derivatives modelled with coefficients of function arctan in (8). There were no approximation of second order derivatives which might explain the comparable lack of accuracy.

The phone recognition results of the CKD3 features, where dynamic parameters were estimated using function (10), were comparable with SPEC and CKD1 features. In turn, we had further explored differences of the SPEC and the CKD3 feature sets in phone classification task. In Fig. 3 is shown a comparison of confusion matrices of SPEC and CKD3.

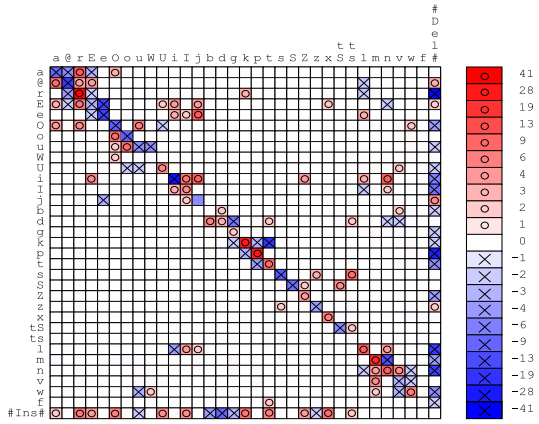

Fig. 3. Subtraction of SPEC and CKD3 confusion matrices shows differences of SPEC (circle) and CKD3 (cross) features in a phone classification task.

A modified confusion matrix [5] in Fig. 3, where the CKD3 confusion matrix was subtracted from the SPEC confusion matrix, shows differences in the phone classification task for both distributions. Modified confusion matrix indicates a better recognition of vowels `/i/`, `/O/`, `/o/`, `/@/` and phone `/s/` with the CKD3 in comparison with the SPEC features. On contrary a better recognition of plosives `/k/`, `/p/`, `/t/`, glides `/l/` and `/r/` and nasal `/m/` could be seen with the SPEC.

There also exists a number of substitutions of the phone /k/ by /t/ and /m/ by /n/ described with the CKD3 features. In addition, more deletions on account of insertions with the CKD3 could be noticed.

5 Conclusion

We proposed a new modified method of speech features extracting based on mel–frequency cepstral coefficients with use of different time–frequency distributions. We also investigated several estimates of the time derivatives obtained from basic measurements of the speech signals. The first and second time derivatives were approximated by regression coefficients, coefficients of function arctan, and functions sine and cosine.

Analyzes and tests are performed for different sets of speech features obtained from spectrogram and cone–shaped kernel distribution using speech recognition system build from hidden Markov models. Effectiveness of the speech features derived from CKD was demonstrated in this study. Coefficients of trigonometric functions were employed in addition to basic cepstral parameters resulting in satisfying phone recognition rate when applied to the Slovene speech database.

References

1. Cohen, L.: Time–Frequency Analysis. Prentince Hall Signal Processing Series (1995)
2. Zhao, Y., Atlas, L. E., Marks, R. J.: The use of cone-shaped kernels for generalized time–frequency representations of nonstationary signals. IEEE Transactions on Acoustics, Speech, Signal Processing, Vol. 38. (1990) 1084–1091
3. Hlawatsch, F., Boudreaux–Bartels, G.F.: Linear and Quadratic Time–Frequency Signal Representations. IEEE SP Magazine, Vol. 9, No. 2. (1992) 21–67
4. Quian, S., Chen, D.: Joint Time–Frequency Analysis: Methods and Applications. Prentince–Hall PTR, New York. (1996)
5. Žibert, J.: Časovno–frekvenčne predstavitve govornih signalov, master thesis. Faculty of Electrical Engineering, University of Ljubljana. (2001)
6. Papoulis, A.: Signal Analysis. McGraw-Hill Book Co., New York. (1996)
7. Loughlin, P.J., Pitton, J.W., Atlas, L.E.: Bilinear Time–Frequency Representations: New Insights and Properties. IEEE Trans. on SP, Vol. 41, No. 2. (1993)
8. Auger, F., Flandrin, P., Gonçalves, P., Lemoine, O.: Time–Frequency Toolbox: For Use with Matlab. Reference Guide. (1996)
9. Picone, J.:Signal Modeling Techniques In Speech Recognition. IEEE Proc. (1993)
10. Furui, S.: Speaker Independent Isolated Word Recogniser Using Dynamic Features of Speech Spectrum. IEEE Transactions on Acoustic, Speech and Signal Processing, Vol. 34, No. 1. (1986) 52–59
11. Dobrišek, S., Mihelič, F., Pavešić, N.: A Multiresolutionally Oriented Approach for Determination of Cepstral Features in Speech Recognition. Proceedings EUROSPEECH'97, Vol. 3. Rhodes, Greece (1997) 1367 – 1370
12. Mihelič, F., Gros, J., Dobrišek, S., Žibert, J., Pavešić, N.: Spoken Language Resources at LAPSC of the University of Ljubljana. Submitted for publication in International Journal of Speech Technology. (2002)
13. Young, S., Odell, J., Ollason, D., Vatchev, V., Woodland, P.: The HTK Book. Cambridge University. Entropic Cambridge Research Laboratory Ltd. (1997)