ON THE USE OF PHASE INFORMATION FOR SPEECH RECOGNITION

Baris Bozkurt and Laurent Couvreur

TCTS Lab, Faculté Polytechnique De Mons, Initialis Scientific Park, B-7000 Mons, Belgium,
phone: +32 65 374733, fax: +32 65 374729,
email: bozkurt,couvreur@tcts.fpms.ac.be, web: http://www.tcts.fpms.ac.be

ABSTRACT

This study addresses the use of short-time phase spectra in automatic speech recognition (ASR). Two recent studies [1,2] have proposed two group delay based spectral representations. Here we propose three new group delay based representations and compare usefulness of all these representations in an ASR experiment. We show that two of the representations we propose perform better, contain equivalent or complementary information to that of the power spectrum and are potentially useful for improving ASR performance.

1. INTRODUCTION

In most state-of-the-art ASR systems, amplitude/power spectrum has been the preferred component of the Fourier Transform (FT) for feature extraction. However, recent studies on speech perception report the importance of phase information [3].

By its nature, the phase component of the FT spectrum is in a wrapped form and the first derivative of the unwrapped phase spectrum, *i.e.* the group delay function, is much easier to study and process. The main difficulties in reliable phase spectrum estimation and unwrapping are mostly related with the zeros of the signal's *z*-transform close to the unit circle, which cause spikes on the derivative of the phase spectrum (group delay) [1]. Methods to remove these spikes are required in order to be able to use the FT phase information in ASR.

Two recent studies address this problem and propose two group delay based features: modified group delay [1] and product spectrum [2]. In this study, we introduce three new group delay based representations and compare all five representations (and the power spectrum for reference) in an ASR experiment. The results show that two of the representations that we propose provide good results and contain equivalent or complementary information to the power spectrum that is potentially useful for improving ASR performance.

The sections are planned as follows: section 2 presents the source of problem in group delay processing, the group delay representations proposed in [1,2] and the representations that we propose in this study. In section 3, we briefly define feature extraction procedures for ASR based on group delay representations (the reader is referred to [2] for more detailed information). Section 4 is dedicated to ASR experiments and finally in section 5, we discuss the results.

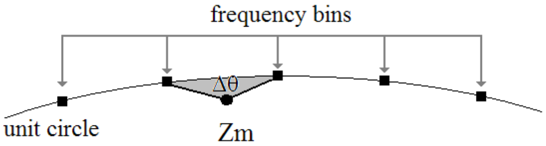

Figure 1: Geometric interpretation of spikes in group delay function at frequency locations close to a root/zero of the *z*-transform polynomial.

2. GROUP DELAY REPRESENTATIONS

2.1 Difficulties In Group Delay Processing

For a given discrete time digital signal *x(n)*, the *z*-transform polynomial, *X(z)*, can be expressed using an all-zero representation as:

$$X(z)=\sum_{n=0}^{N-1}x(n)z^{-n}=x(0)z^{-N+1}\prod_{m=1}^{N-1}(z-Z_m)$$

where Z_m are the roots of the *z*-transform polynomial. The FT, which is simply the *z*-transform computed on the unit circle, can be expressed as:

$$X(\omega)=x(0)\left(\rho e^{j\theta(\omega)}\right)^{(-N+1)}\prod_{m=1}^{N-1}(\rho e^{j\theta(\omega)}-Z_m) \quad (1)$$

where the radius $\rho=1$. Each factor in Eq. 1 corresponds, in the *z*-plane, to a vector starting at Z_m and ending at $e^{j\theta(\omega)}$. As illustrated in Fig. 1, the group delay, *i.e.* the rate of change in the phase component, is very high at frequency bins very close to a root/zero of the *z*-transform polynomial and becomes ill-defined when the zero coincides with a frequency bin.

For actual windowed speech signals, many zeros appear to be very close to the unit circle. The effect of zeros close to the unit circle can easily be observed both on amplitude spectra (as dips) and group delay functions (as large spikes). For the amplitude spectrum a spectral envelope with formant peaks is still observed/available with the presence of dips due to the zeros. Unlike amplitude spectrum, the group delay function is effected to an extend that the spikes hide the actual vocal tract phase/group delay information (see Fig. 2 for an example). Therefore, it is necessary to find means of avoiding such spikes in the group delay function for using it in feature extraction for ASR systems. Below, we present two methods from recent literature followed by our three methods for the estimation of group delay representations, which can be further processed for feature extraction.

2.2 Modified Group Delay Function (MODGDF) [1]

The group delay function can be expressed as:

$$\tau_p(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \tag{2}$$

where $X(\omega)$ and $Y(\omega)$ denote to the FT of $x(n)$ and $nx(n)$ respectively, R and I refer to real and imaginary parts [2]. This formulation is quite useful since there is no need for phase unwrapping, which is often referred to be problematic, and group delay can be directly computed using the FT transform only. In [1], the authors propose the so-called modified group delay function (MODGDF), which is a modified version of Eq. 2:

$$\tau_{\mathrm{mod}}(\omega) = \left(\frac{\tau_p(\omega)}{|\tau_p(\omega)|} \right) \left(|\tau_p(\omega)| \right)^{\alpha}$$

$$\tau_p(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{S(\omega)^{2\gamma}}$$

where the term $|X(\omega)|$ is replaced by its cepstrally smoothed version $S(\omega)$ in order to reduce spikes on the group delay function. This is because the term $|X(\omega)|$ in the denominator in Eq. 2 gets very small when there exists a zero very close to the unit circle. In addition, two new parameters are introduced: α and γ which need to be fine-tuned according to the environment. In all tests/plots of this study, we have set the parameters as in [1], namely α=0.4 and γ=0.9. These smoothing parameters also reduce the effect of the spikes on the group delay function to some extend.

2.3 Product Spectrum (PS) [2]

In [2], another group delay based representation is proposed. It is a version of Eq. 2 where the denominator, which is considered to be the source of spikes, is removed. The product spectrum $Q(\omega)$ is defined as the product of the power spectrum and the group delay function:

$$Q(\omega) = |X(\omega)|^2 \tau_p(\omega) = X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)$$

2.4 Group Delay of GCI-Synchronously Windowed Speech (GDGCI)

Recently, we have shown that smooth group delay functions can be directly computed from speech signals if windowing is appropriately [4]: window centred at glottal closure instant (GCI) and smaller than two pitch periods. In our computations, we use a Blackman window function though other types are possible, *e.g.* Gaussian or Hanning-Poisson. Such a windowing operation results in grouping the zeros across the unit circle but not on it, therefore avoiding most of the spikes. GDGCI is referring to the group delay function computed using Eq. 2 on GCI-synchronously windowed speech data. GCI detection is achieved by processing the centre-of-gravity evolution signal obtained by shifting an analysis window on the speech signal as described in [5].

2.5 Chirp Group Delay of GCI-Synchronously Windowed Speech (CGDGCI)

After tests on noisy real speech data, we have observed that GCI-synchronous windowing cannot guarantee a completely zero-free region on the unit circle. The zeros related to the noise signal component may appear on the unit circle and introduce extra spikes. In addition, the size of window appears to be a problem for very high pitch speech: an error to include more than two pitch periods in the speech frame results in zeros due to periodicity on the unit circle. For this reason a two-step method is proposed: suppression of zeros outside the unit circle (which are mainly due to the glottal flow component of speech [6]) and computation of the phase derivative on a circle outside the unit circle from the remaining zeros using Eq. 1.

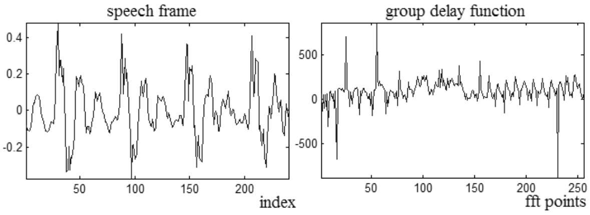

Figure 2: Time-domain signal of a 30 ms speech frame and its group delay function. The frame example is extracted from the noise-free utterance “mah_4625" of the test set A of the AURORA-2 [8] and corresponds to vowel /i/ in word “6”.

The roots of a high degree polynomial can be efficiently obtained by searching for the eigenvalues of the associated companion matrix [7]. The procedure provides enough accuracy to carry reliable spectral analysis. Unfortunately, such algorithm for estimating polynomial roots is computationally demanding.

In order to obtain a very smooth group delay function with well-resolved formant peaks, it is useful to compute the group delay function on a circle other than the unit circle in the z-plane. Equivalently, one can compute the group delay function of a chirped version of the signal, *i.e.* multiplied by a decaying exponential whose damping parameter is solely related to the actual circle radius ρ. It results in the GCI-synchronous chirped group delay function (CGDGCI). In [6], we have shown that vocal tract formants can be estimated easily by tracking the peaks of CGDGCI. The choice of the radius ρ of the z-transform computation circle is a compromise between having smooth/blurred or detailed/spiky spectral representation. The value ρ=1.12 is observed to be a good choice.

2.6 Chirp Group Delay of The Zero-Phase Version (CGDZP)

Finally we propose another group delay function, for which heavy computation of polynomial roots is not necessary. Again the procedure contains two steps: computation of the zero-phase version of the signal (inverse FT of $|X(\omega)|$) and computation of the CGD on the circle with ρ=1.12 using the chirp z-transform.

Conversion to zero-phase guarantees that all of the zeros occur very close to the unit circle therefore the resulting chirp group delay representation is very smooth with well-resolved formant peaks. However, the phase information is destroyed for this case, therefore the representation contains only the information available in the amplitude spectrum but formant peak resolutions appear with higher resolution.

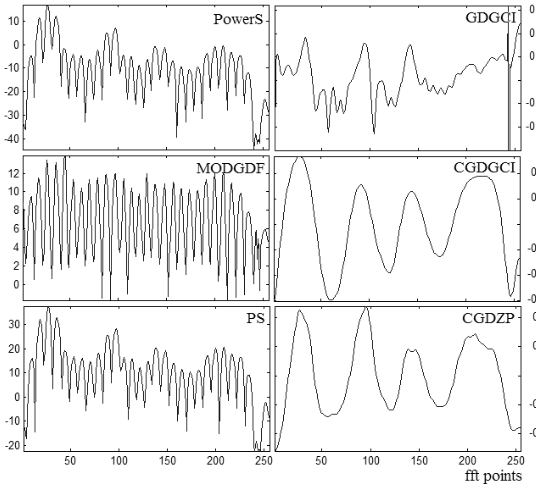

Figure 3: Power spectrum (PowerS) and group delay representations for the speech signal frame in Fig. 2.

2.7 Comparison of Proposed Methods via Spectral Plots

Fig. 2 presents a typical time-domain speech signal and its group delay function. As expected, the group delay function computed directly on the speech frame contains mainly spikes and resonance information cannot be observed. In Fig. 3, we present the five group delay based representations together with the power spectrum for this speech frame. The formant peaks appear with high resolution in GDGCI, CGDGCI and CGDZP. GDGCI includes a spike at high frequencies due to a zero, which cannot be avoided by only GCI-synchronous windowing. As more noise was added to signals, such spikes would be more frequent, therefore the robustness of GDGCI to noise is rather low. Thanks to zero removal techniques and zero-phasing, CGDGCI and CGDZP are more robust to noise.

In Fig. 4, we also present spectrogram plots obtained using the described group delay representations as well as the classical power spectrum. The formant tracks can be well observed on all of the spectrograms except for MODGDF, and PS is very close to PowerS as already shown in Fig. 1 of [3] and in Fig. 3 above. GDGCI representation is vague to some level. This is mainly due to the fact that unvoiced frames include spikes with large amplitudes that force a low contrast on the plots. Actually, the group delay functions computed on unvoiced frames mostly do not contain resonance information but random spikes. GDGCI and CGDGCI are actually the two representations that really suffer from this problem.

These observations suggest that the representations have some potential in an ASR framework. The main concern is if they can provide complementary information to the power spectrum and improve performance.

3. COMPUTATION OF FEATURES FOR ASR

The most common feature extraction for ASR systems consists of computing power-based Mel-frequency cepstral coefficients (MFCC) [9], that is, a Mel filterbank is applied to the power spectrum and an inverse discrete cosine transform (IDCT) is computed on the logarithm of its outputs. The main reason for such processing is to capture the essential shape of the power spectrum with a few coefficients well conditioned for pattern recognition. A similar scheme can be applied to the group delay functions in order to derive phase-based feature extractions for ASR systems. The simplest approach consists in replacing the power spectrum in the MFCC algorithm by the group delay representation computed via one of the analysis techniques described in the previous section.

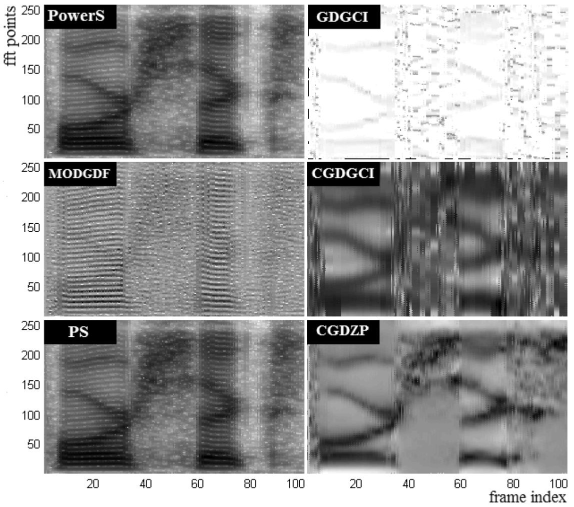

Figure 4: Spectrogram plots of the noise-free utterance “mah_4625a”. Only the first half of the signal that to the digit utterance “46” is presented.

In this work, we use a Mel filterbank with 24 triangular filters and 12 IDCT coefficients are computed for 30 ms frames shifted by 10 ms. Note that the logarithm is not applied on the outputs of the filterbank when fed with a phase spectrum. These coefficients are augmented with the frame log-energy and their (delta-)delta coefficients. We finally come up with six feature extractions: MFCC as a reference and five group delay based methods.

4. ASR EXPERIMENTS

4.1 ASR system

The ASR system that is considered in this work relies on the STRUT toolkit [10]. It merely consists of three blocks. First, the feature extraction chops the discrete speech signal into overlapping frames and computes for each frame a set of acoustic coefficients using one of the algorithms described in the previous sections. Next, the acoustic coefficient vectors are fed into the acoustic model that is here based on the Multi Layer Perceptron (MLP) / Hidden Markov Models (HMM) paradigm [11]. In this framework, the phonemes of the language under consideration are modelled by HMM's whose observation state probabilities are estimated as the outputs of a MLP. Such an acoustic model is trained beforehand in a supervised fashion on a large speech database containing a few hours of phonetically segmented speech material. Finally, the word decoder searches for the most likely word sequence given the sequence of probability vectors for all the frames. Here, the search is constrained by a phonetic lexicon

Table 1: Performances of ASR system for various feature extraction on the AURORA-2 task. Results are given in terms of word error rate (WER) in percent. For every SNR value, the rightmost value corresponds to the case where the actual feature extraction is combined with MFCC as described in section 4.3.

Feature Extraction	SNR (dB)													
	∞		20		15		10		5		0		-5	
MFCC	1.9	–	6.7	–	18.6	–	45.2	–	75.1	–	88.8	–	91.5	–
MODGDF	3.2	2.1	12.3	8.5	25.6	23.9	50.8	52.7	80.8	79.5	97.1	89.5	99.8	91.5
PS	2.0	1.9	6.7	6.7	19.4	18.6	45.3	44.4	75.5	74.6	89.0	88.5	92.2	91.6
GDGCI	8.8	2.1	32.8	7.8	49.4	16.8	69.0	36.0	88.3	64.4	98.6	88.0	100.0	96.1
CGDGCI	3.2	1.8	12.3	5.8	25.6	12.2	50.8	29.1	80.8	58.0	97.0	83.8	99.8	93.8
CGDZP	1.8	1.7	5.8	5.0	12.2	10.4	29.4	24.8	62.6	52.7	88.7	82.3	97.6	91.1

and a word grammar, which together define all the authorized sequences of phonemes. Here, the search is performed as a one-pass frame-synchronous Viterbi algorithm [9] without any pruning constraints.

4.2 Speech Database

The AURORA-2 database [8] was used in this work. It consists of connected English digit utterances sampled at 8kHz. More exactly, we used the clean training set, which contains 8440 noise-free utterances spoken by 110 male and female speakers, for building our acoustic models. These models were evaluated on the test set A. It has 4004 different noise-free utterances spoken by 104 other speakers. It also contains the same utterances corrupted by four types of real-world noises (subway, babble, car, exhibition hall) at various signal-to-noise ratios (SNR) ranging from 20dB to -5dB. During the recognition experiments, the decoder is constrained by a lexicon reduced to the English digits and no grammar is applied.

4.3 Experimental Results

Tab. 1 gives the word error rates (WER) for the ASR system tested with the feature extractions described in section 3. Errors are counted in terms of word substitutions, deletions and insertions, and error rates are averaged over all noise types. The results are also provided when combining MFCC feature extraction with the others (rightmost figure for every noise level and feature extraction). The combination is simply performed by taking a weighted geometric average of the probability outputs of the combined acoustic models:

$$p_{12} = p_1^{\lambda} \cdot p_2^{1-\lambda}$$

where p_{12}, p_1 and p_2 denote the combined probability and the probability provided by the two combined acoustic model, respectively. The combination parameter λ takes its value in the range (0,1) and is optimised for every combination.

5. DISCUSSIONS AND CONCLUSIONS

Our main target in this study is to test if a phase/group delay representation carries equivalent or complementary information to that of the power spectrum in the framework of feature extraction for ASR systems. The results presented in Tab. 1 shows that the group delay representations CGDGCI and CGDZP have this potential: the rightmost values compared to the MFCC-only results are in all cases lower except for the extreme noise setting SNR=-5dB.

In our in-detailed analysis, we have observed that the GDGCI, which is the pure group delay function computed on GCI-synchronous data without further processing, mainly suffers from window size problems (including several pitch periods result in zeros on the unit circle). In addition, GDGCI and CGDGCI do not carry reliable information for unvoiced frames.

The AURORA-2 task was chosen for its simplicity and ease of comparison to the already available results in [2]. Further experiments will be performed on other tasks in order to confirm the present results about the usefulness of phase information for ASR systems.

REFERENCES

[1] R. M. Hegde, H. A. Murthy and V. R. R. Gadde, "Continuous speech recognition using joint features derived from the modified group delay function and MFCC", *in Proc. ICSLP*, Jeju, Korea, Oct. 2004.

[2] D. Zhu and K. K. Paliwal, "Product of power spectrum and group delay function for speech recognition", in *Proc. ICASSP*, Montreal, Canada, May 2004.

[3] K. K. Paliwal and L. D. Alsteris, "Usefulness of phase spectrum in human speech perception", in *Proc. EUROSPEECH*, Geneva, Switzerland, Sep. 2003.

[4] B. Bozkurt, B. Doval, C. D'Alessandro and T. Dutoit, "Appropriate windowing for group delay analysis and roots of *z*-transform of speech signals", in *Proc. EUSIPCO*, Vienna, Austria, Sep. 2004.

[5] H. Kawahara, Y. Atake, and P. Zolfaghari, "Accurate vocal event detection method based on a fixed-point to weighted average group delay", in *Proc. ICSLP*, Beijing, China, Oct. 2000.

[6] B. Bozkurt, B. Doval, C. D'Alessandro and T. Dutoit, "Improved differential phase spectrum processing for formant tracking", in *Proc. ICSLP*, Jeju, Korea, Oct. 2004.

[7] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd Edition, Johns Hopkins University Press, 1996.

[8] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition Systems under noisy conditions", in *Proc. ASR2000*, Paris, France, Sep. 2000.

[9] X. Huang, A. Acero and H.-W. Hon, *Spoken Language Processing: A guide to Theory, Algorithm, and System Development*, Prentice Hall, 2001.

[10] J.-M. Boite, L. Couvreur, S. Dupont and C. Ris, *Speech Training and Recognition Unified Tool (STRUT)*, http://tcts.fpms.ac.be/asr/project/strut.

[11] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*, Kluwer Academic Publisher, 1994.