INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Harmonic-aligned Frame Mask Based on Non-stationary Gabor Transform with Application to Content-dependent Speaker Comparison

Feng Huang, Peter Balazs

Acoustic Research Institute, Austrian Academy of Sciences

fhuang@kfs.oeaw.ac.at, peter.balazs@oeaw.ac.at

Abstract

We propose harmonic-aligned frame mask for speech signals using non-stationary Gabor transform (NSGT). A frame mask operates on the transfer coefficients of a signal and consequently converts the signal into a counterpart signal. It depicts the difference between the two signals. In preceding studies, frame masks based on regular Gabor transform were applied to single-note instrumental sound analysis. This study extends the frame mask approach to speech signals. For voiced speech, the fundamental frequency is usually changing consecutively over time. We employ NSGT with pitch-dependent and therefore time-varying frequency resolution to attain harmonic alignment in the transform domain and hence yield harmonic-aligned frame masks for speech signals. We propose to apply the harmonic-aligned frame mask to content-dependent speaker comparison. Frame masks, computed from voiced signals of a same vowel but from different speakers, were utilized as similarity measures to compare and distinguish the speaker identities (SID). Results obtained with deep neural networks demonstrate that the proposed frame mask is valid in representing speaker characteristics and shows a potential for SID applications in limited data scenarios.

Index Terms: Non-stationary Gabor transform, frame mask, harmonic alignment, pitch-dependent frequency resolution, speaker feature, speaker comparison

1. Introduction

Time-frequency (TF) analysis is the foundation of audio and speech signal processing. The short-time Fourier transform (STFT) is a widely used tool, which can be effectively implemented by FFT [1]. STFT features straightforward interpretation of a signal. It provides uniform time and frequency resolution with linearly-spaced TF bins. The corresponding theory was generalized in the framework of Gabor analysis and Gabor frames [2, 3, 4].

Signal synthesis is an important application area of time-frequency transforms. Signal modification, denoising, separation and so on can be achieved by manipulating the analysis coefficients to synthesize a desired one. The theory of Gabor multiplier [5] or, in general terms, frame multiplier [6, 7] provides a basis for the stability and invertibility of such operations. A *frame multiplier* is an operator that converts a signal into another by pointwise multiplication in the transform domain for resynthesis. The sequence of multiplication coefficients is called a *frame mask* (or symbol). Such operators allow easy implementation of time-varying filters [8]. They have been used in perceptual sparsity [9], denoising [10] and signal synthesis [11]. Algorithms to estimate frame mask between audio signals were investigated in [11, 12], where it was demonstrated that the frame mask between two instrumental sounds (of a same note) was an effective measure to characterize timbre variations between the instruments. Such masks were used for timbre morphing and instrument categorization. In this paradigm, the two signals were of the same fundamental frequency and their harmonics were naturally aligned, which vouched for the prominence of the obtained mask for TF analysis/synthesis with uniform resolution.

This study extends the frame mask method to speech signals. One intrinsic property of (voiced) speech signal is that the fundamental frequency (f_0 or pitch) varies consecutively over time. Therefore, the harmonic structures are not well aligned when comparing two signals. We propose to employ the non-stationary Gabor transform (NSGT) [13] to tackle this issue. NSGT provides flexible time-frequency resolution by incorporating dynamic time/frequency hop-size and dynamic analysis windows [13, 14, 15]. We develop an NSGT whose frequency resolution changes over time. We set the frequency hop-size in ratio to f_0 to achieve harmonic alignment (or partial alignment cf. Section 4) in the transform domain. On this basis, we propose the harmonic-aligned frame mask. To demonstrate feasibility in speech, we shall evaluate the proposal in the context of vowel-dependent speaker comparison. Frame marks between voiced signals of the same vowel but pronounced by different speakers are proposed as similarity measures for speaker characteristics to distinguish speaker identities in a limited data scenario (cf. Section 5 for details).

This paper is organized as follows. In Section 2, we briefly review frame and Gabor theory. In Section 3, we elaborate frame mask and the previous application in instrumental sound analysis. In Section 4, we develop the non-stationary Gabor transform with pitch-dependent frequency resolution and propose the harmonic-aligned frame mask. Section 5 presents the evaluation in vowel-dependent speaker identification. And finally, Section 6 concludes this study.

2. Preliminaries and notation

2.1. Frame theory

Denote by $\{g_\lambda : \lambda \in \Lambda\}$ a sequence of signal atoms in the Hilbert space $\mathcal{H}$, where Λ is a set of index. This atom sequence is a frame [3] if and only if there exist constants A and B, $0 < \mathsf{A} \leq \mathsf{B} < \infty$, such that

$$\mathsf{A}\|f\|_2^2 \leq \sum_\lambda |c_\lambda|^2 \leq \mathsf{B}\|f\|_2^2,\ \forall f \in \mathcal{H}. \tag{1}$$

where $c_\lambda = \langle f, g_\lambda \rangle$ are the *analysis coefficients*. A and B is called the lower and upper frame bounds, respectively. The *frame operator* $\mathbf{S}$ is defined by $\mathbf{S}f = \sum_\lambda \langle f, g_\lambda \rangle g_\lambda$.

Given $\{h_\lambda = S^{-1} g_\lambda : \lambda \in \Lambda\}$ the canonical dual frame of $\{g_\lambda : \lambda \in \Lambda\}$, f can be perfectly reconstructed from the analysis coefficients by

$$f = \sum_\lambda \langle f, g_\lambda \rangle h_\lambda. \tag{2}$$

http://dx.doi.org/10.21437/Interspeech.2019-1327

The dual frame always exists [16], and for redundant cases there are infinitely many other duals allowing reconstruction.

2.2. Discrete Gabor transform

We take the Hilbert space $\mathcal{H}$ to be $\mathbb{C}^L$. Given non-zero prototype window $g = (g[0], g[1], \cdots, g[L-1])^T \in \mathbb{C}^L$, the translation operation $\mathbf{T}_x$ and modulation operator $\mathbf{M}_y$ are, respectively, defined as

$$\mathbf{T}_x g[l] = g[l-x] \text{ and } \mathbf{M}_y g[l] = g[l] e^{\frac{2\pi i y l}{L}},$$

where $x, y \in \mathbb{Z}_L$ and the translation is performed modulo L. For selected constants $\mathsf{a}, \mathsf{b} \in \mathbb{Z}_L$, with some $N, M \in \mathbb{N}$ such that $N\mathsf{a} = M\mathsf{b} = L$, we take Λ to be a regular discrete lattice, i.e., $\lambda = (m, n)$, and obtain the *Gabor system* [2] $\{g_{m,n}\}_{m \in \mathbb{Z}_M, n \in \mathbb{Z}_N}$ as

$$g_{m,n}[l] = \mathbf{T}_{n\mathsf{a}} \mathbf{M}_{m\mathsf{b}} g[l] = g[l - n\mathsf{a}] e^{\frac{2\pi i m \mathsf{b}(l - n\mathsf{a})}{L}}. \quad (3)$$

If $\{g_{m,n}\}_{m,n}$ satisfies (1) for $\forall f \in \mathbb{C}^L$, it is called a Gabor frame [17]. The discrete Gabor transform (DGT) of $f \in \mathbb{C}^L$ is a matrix $C = \{c_{m,n}\} \in \mathbb{C}^{M \times N}$ with $c_{m,n} = \langle f, g_{m,n} \rangle$. The associated frame operator $\mathbf{S} : \mathbb{C}^L \to \mathbb{C}^L$ reads

$$\mathbf{S} f = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \langle f, g_{m,n} \rangle g_{m,n}. \quad (4)$$

The canonical dual frame $\{\widetilde{g}_{m,n}\}_{m,n}$ of the Gabor frame $\{g_{m,n}\}_{m,n}$ is given by $\widetilde{g}_{m,n} = \mathbf{T}_{n\mathsf{a}} \mathbf{M}_{m\mathsf{b}} \mathbf{S}^{-1} g$ [18], with which f can be perfectly reconstructed by

$$f = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} c_{m,n} \widetilde{g}_{m,n}.$$

Note that the DGT coefficients are essentially sampling points of the STFT of f with window g at the time-frequency points $(n\mathsf{a}, m\mathsf{b})$, with a and b being the sampling step (i.e., hop-size) in time and frequency [18]. In non-stationary settings, the hop-sizes are allowed to be variant (cf. Section 4).

3. Frame mask for instrumental sound analysis

3.1. Frame mask

Consider a pair of frames $\{g_\lambda : \lambda \in \Lambda\}$ and $\{h_\lambda : \lambda \in \Lambda\}$. A *frame multiplier* [19], denoted by $\mathbb{M}_{\boldsymbol{\sigma}; g, h}$, is an operator that acts on a signal by pointwise multiplication in the transform domain. The symbol $\boldsymbol{\sigma} = \{\sigma_\lambda, \lambda \in \Lambda\}$ is a sequence that denotes the multiplication coefficients. For signal f

$$\mathbb{M}_{\boldsymbol{\sigma}; g, h} f = \sum_\lambda \sigma_\lambda \langle f, g_\lambda \rangle h_\lambda. \quad (5)$$

Here $\boldsymbol{\sigma}$ is called a *frame mask*. In the considered signal analysis/transform domain, $\boldsymbol{\sigma}$ can be viewed as a *transfer function*.

When Gabor frames $\{g_{m,n}\}_{m,n}$ and $\{h_{m,n}\}_{m,n}$ are considered, we set $\lambda = (m, n)$. In this case the frame multiplier in (5) is known as Gabor multiplier. The corresponding frame mask $\boldsymbol{\sigma} = \{\sigma_{m,n}\} \in \mathbb{C}^{M \times N}$ is also known as *Gabor mask*.

3.2. For instrument timbre analysis and conversion

The application of frame masks in musical signals was investigated in [11, 12]. Based on DGT, the proposed signal model converts one sound into another by

$$f^{\mathrm{B}} = \mathbb{M}_{\boldsymbol{\sigma}^{\overrightarrow{\mathrm{AB}}}; g, \widetilde{g}} f^{\mathrm{A}} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \sigma_{m,n}^{\overrightarrow{\mathrm{AB}}} \langle f^{\mathrm{A}}, g_{m,n} \rangle \widetilde{g}_{m,n}, \quad (6)$$

where $f^{\mathrm{A}}, f^{\mathrm{B}} \in \mathbb{R}^L$ are two audio signals and $\boldsymbol{\sigma}^{\overrightarrow{\mathrm{AB}}}$ is the unknown mask to be estimated. An obvious solution is to set $\sigma_{m,n}^{\overrightarrow{\mathrm{AB}}} = c_{m,n}^{\mathrm{B}} / c_{m,n}^{\mathrm{A}}$, where $c_{m,n}^{\mathrm{A}}$ and $c_{m,n}^{\mathrm{B}}$ are the DGT coefficients of f^{A} and f^{B}, respectively. However, this solution is non-stable and unbounded as the DGT coefficients in the denominator can be 0 or very small. To guarantee existence of a stable solution, it was proposed to estimate the mask via

$$\min_{\boldsymbol{\sigma}^{\overrightarrow{\mathrm{AB}}}} \| f^{\mathrm{B}} - \mathbb{M}_{\boldsymbol{\sigma}^{\overrightarrow{\mathrm{AB}}}; g, \widetilde{g}} f^{\mathrm{A}} \|^2 + \mu d(\boldsymbol{\sigma}^{\overrightarrow{\mathrm{AB}}}), \quad (7)$$

with a (convex) regularization term $d(\boldsymbol{\sigma}^{\overrightarrow{\mathrm{AB}}})$, whose influence is controlled by the parameter μ [12]. As the existence of a stable solution is assured, such approach in general can be applied to arbitrary pair of signals. However, it might be difficult to interpret the estimated masks (e.g., the mask between two pure-tone signals with different fundamental frequencies).

Given that f^{A} and f^{B} are of the same note produced by different instruments, the frame mask between the two signals was found to be effective to characterize the timbre difference between the two instruments [11, 12]. Such masks were utilized as similarity measures for instrument classification and for timbre morphing and conversion. Rationality of these applications roots from two aspects:

1) Instrumental signals of a same note possess the same fundamental frequency. Harmonic structures of the signals are naturally aligned.
2) DGT performs TF analysis over a regular TF lattice, and consequently preserves the property of harmonic alignment in the transform domain.

4. Frame mask for speech signals using non-stationary Gabor transform

Similar to audio sounds of instrument notes, (voiced) speech signals are also harmonic signals. Analog to the above-mentioned applications, this study explores the application of frame mask in speech signals. In particular, we consider to use voiced speech as source and target signals and to estimate the frame mask between them. We are specially interested in the case that the source and the target are of the same content, e.g., the same vowel. For such a case, a valid frame mask could measure specific variations among the signals, such as speaker variations.

Nevertheless, attempting to use (7) for speech signals, we immediately face a fundamental problem. For speech signals, the fundamental frequency usually varies over time consecutively. Therefore, harmonic structures of the source and target voice are mostly not aligned. To address this problem, we propose to employ non-stationary Gabor transform, which allows flexible time-frequency resolution [13]. Within the framework of non-stationary Gabor analysis, we intend to achieve dynamic alignment of the signals' harmonic structures. In the following, we shall develop NSGT with pitch-dependent frequency resolution to achieve harmonic alignment in the transform domain, and shall propose the harmonic-aligned frame mask for speech signals on that basis.

4.1. Non-stationary Gabor transform with pitch-dependent frequency resolution

We consider analyzing a voiced signal $f \in \mathbb{R}^L$ with a window g that is symmetric around zero. As the stationary case in Section 2.2, we use a constant time hop-size a, resulting in $N = \frac{L}{\mathsf{a}} \in \mathbb{N}$ sampling points in time for the TF analysis. However, we set the frequency hop-size according to the fundamental frequency of the signal (see Remark 2.1 for discussion on pitch estimation issue). Following the quasi-stationary assumption for speech signals, we assume that the fundamental frequency is approximately fixed within the interval of the analysis window. At time n, let $f_0(n\mathsf{a})$ denote the fundamental frequency in Hz, we set the corresponding frequency hop-size as

$$b_n^{f_0} = \left\lfloor \frac{\mathsf{p} f_0(n\mathsf{a})}{\mathsf{q}} \Big/ \frac{f_s}{L} \right\rceil, \tag{8}$$

where $\mathsf{p}, \mathsf{q} \in \mathbb{N}$ are a pair of parameters to be set. $\lfloor \rceil$ denotes rounding to the closest positive integer, and f_s is the signal's sampling rate in Hz. With (8), q frequency sampling points are deployed per $\mathsf{p} f_0(n\mathsf{a})$ Hz. The total number of frequency sampling points at n is hence $M_n^{f_0} = L / b_n^{f_0} \in \mathbb{N}$. Consequently, we obtain the pitch-depenent *non-stationary Gabor system* (NSGS) $\{g_{m,n}\}_{m \in \mathbb{Z}_{M_n^{f_0}}, n \in \mathbb{Z}_N}$ as

$$g_{m,n}[l] = \mathbf{T}_{n\mathsf{a}} \mathbf{M}_{m b_n^{f_0}} g[l] = g[l - n\mathsf{a}] e^{\frac{2\pi i m b_n^{f_0}(l - n\mathsf{a})}{L}}. \tag{9}$$

It is called a *non-stationary Gabor frame* (NSGF) if it fulfills (1) for $\mathbb{C}^L$. The sequence $\{c_{m,n}\}_{m,n} = \{\langle f, g_{m,n} \rangle\}_{m,n}$ are the *non-stationary Gabor transform* coefficients. In general, due to the dynamic frequency hop-size, these coefficients do not form a matrix.

Eq. (8) features a time-varying and pitch-dependent frequency resolution. More importantly, it allows harmonic alignment in the NSGT coefficients with respect to the frequency index m. For example, with $\mathsf{p} = 1$, for any n, $c_{\mathsf{q},n}, c_{2\mathsf{q},n}, c_{3\mathsf{q},n}, \cdots$ naturally correspond to the harmonic frequencies of the signal. The parameter p allows performing partial alignment wrt. integer multiples of the p-th harmonic frequency.

Remark 1. To satisfy $M_n^{f_0} b_n^{f_0} = L, \forall n \in \mathbb{Z}_N$, zero-padding for f may be needed for an appropriate L. If an extremely large L is required, it is always practicable to divide the signal into segments of shorter duration using overlap-and-add windows, and obtain NSGT coefficients for each segment separately. A practical example for such procedure can be found in [14].

Now we consider the canonical dual $\{\widetilde{g}_{m,n}\}_{m,n}$. Denote by $\mathrm{supp}(g) \subseteq [\mathsf{c}, \mathsf{d}]$ the support of the window g, i.e., the interval where the window is nonzero. We choose $M_n^{f_0} \geq \mathsf{d} - \mathsf{c}, \forall n \in \mathbb{Z}_N$, which is referred to as the *painless* case [13]. In other words, we require the frequency sampling points to be dense enough. In this painless case, we have the following [13].

Proposition 1. If $\{g_{m,n}\}_{m,n}$ is a painless-case NSGF, then the frame operator $\mathbf{S}$ (cf. (4)) is an $L \times L$ diagonal matrix with diagonal element

$$s_{l,l} = \sum_{n=0}^{N-1} M_n^{f_0} \left| g[l - n\mathsf{a}] \right|^2 > 0, \forall l \in \mathbb{Z}_L. \tag{10}$$

And the canonical dual frame $\{\widetilde{g}_{m,n}\}_{m,n}$ is given by

$$\widetilde{g}_{m,n}[l] = \frac{g_{m,n}[l]}{s_{l,l}}. \tag{11}$$

4.2. Harmonic-aligned frame mask

In this section, we present a general form of frame mask based on the above pitch-dependent NGST. For two voiced signals $f^{\mathrm{A}}, f^{\mathrm{B}} \in \mathbb{R}^L$, denote their fundamental frequency by $f_{\mathrm{A};0}$ and $f_{\mathrm{B};0}$, respectively. Using (9) with the same window g and the same time hop-size a for both signals, we construct two Gabor systems $\{g_{m,n}^{\mathrm{A}}\}_{m \in \mathbb{Z}_{M_n^{f_{\mathrm{A};0}}}, n \in \mathbb{Z}_N}$ and $\{g_{m,n}^{\mathrm{B}}\}_{m \in \mathbb{Z}_{M_n^{f_{\mathrm{B};0}}}, n \in \mathbb{Z}_N}$. Denote $\overleftrightarrow{M} = \max\left(\max_n(M_n^{f_{\mathrm{A};0}}), \max_n(M_n^{f_{\mathrm{B};0}})\right)$. To simplify the presentation of the concept without losing the frame property (1), we can consider extend the two systems as $\{g_{m,n}^{\mathrm{A}}\}_{m \in \mathbb{Z}_{\overleftrightarrow{M}}, n \in \mathbb{Z}_N}$ and $\{g_{m,n}^{\mathrm{B}}\}_{m \in \mathbb{Z}_{\overleftrightarrow{M}}, n \in \mathbb{Z}_N}$ e.g., with periodic extension to the modulation operator wrt. the index m. Under such circumstance, we can denote the NGST coefficients in matrix forms as $C^{\mathrm{A}} = \{c_{m,n}^{\mathrm{A}}\}_{m,n} \in \mathbb{C}^{\overleftrightarrow{M} \times N}$ and $C^{\mathrm{B}} = \{c_{m,n}^{\mathrm{B}}\}_{m,n} \in \mathbb{C}^{\overleftrightarrow{M} \times N}$. The *harmonic-aligned frame mask* (HAFM) $\boldsymbol{\sigma}^{\overrightarrow{\mathrm{AB}}} \in \mathbb{C}^{\overleftrightarrow{M} \times N}$ between the two voiced signals therefore acts as

$$f^{\mathrm{B}} = \mathbb{M}_{\boldsymbol{\sigma}^{\overrightarrow{\mathrm{AB}}}; g^{\mathrm{A}}, \widetilde{g}^{\mathrm{B}}} f^{\mathrm{A}} = \sum_{m=0}^{\overleftrightarrow{M}-1} \sum_{n=0}^{N-1} \sigma_{m,n}^{\overrightarrow{\mathrm{AB}}} \langle f^{\mathrm{A}}, g_{m,n}^{\mathrm{A}} \rangle \widetilde{g}_{m,n}^{\mathrm{B}}. \tag{12}$$

To estimate the frame mask, existing methods [11, 12] for the problem in (6) can be directly applied. For both Gabor systems $\{g^{\mathrm{A}}\}$ and $\{g^{\mathrm{B}}\}$, the parameters p and q in (8) need to be appropriately set. We consider to set q for both systems to the same value. However, depending on specifics of the source and target signal (as well as the application purpose), the parameter p may be set to different values for both systems. *Example* 1: If $f_{\mathrm{A};0}$ and $f_{\mathrm{B};0}$ are close (enough), we consider $\mathsf{p} = 1$ for both Gabor systems. This leads to a one-to-one alignment of all harmonics. *Example* 2: If $f_{\mathrm{A};0}$ and $f_{\mathrm{B};0}$ are significantly different in value, we may consider an anchor frequency F and set $\mathsf{p}^{\mathsf{A}} = \lfloor F / f_{\mathrm{A};0} \rceil$, $\mathsf{p}^{\mathsf{B}} = \lfloor F / f_{\mathrm{B};0} \rceil$. This results in partial alignment of the harmonics, i.e., only the harmonics around F and its multiples are aligned.

Remark 2. 1) The proposed approach practically depends on a reliable method to estimate the fundamental frequencies. A thorough discussion of such topic is beyond the scope of this paper. In the evaluation, we applied the methods in [20, 21].
2) It may be a false impression that pitch independence is achieved in the frame masks by the harmonic alignment. On the contrary, the resulted frame mask is essentially dependent on the fundamental frequencies. It equivalently describes the variations between two spectra which are warped in a pitch-dependent and linear way. It contains information related to the spectral envelopes and also highly depends on the fundamental frequencies. It is our interests to utilize the proposed mask as feature measure for classification tasks.

5. Evaluation in content-dependent speaker comparison

We now evaluate harmonic-aligned frame masks for speaker identity comparison in a content-dependent context. In particular, the source and target signals are of the some vowel but pronounced by different speakers. In this setting, we estimate the frame masks between an input speaker and a fixed reference speaker. For different speakers, we compare them to the same reference speaker, and use the estimated masks as speaker feature to measure and distinguish the speaker identities. It can be considered as a task of close-set speaker identification with content-dependent and limited-data constraints (see the experimental settings in 5.1).

To estimate the harmonic-aligned frame mask, we adopt the approach (7) and use transform domain proxy [11]. For our case, the first item in (7) can be written as $\left\| \sum_m \sum_n (c_{m,n}^{\mathrm{B}} - \sigma_{m,n}^{\overrightarrow{\mathrm{AB}}} c_{m,n}^{\mathrm{A}}) \widetilde{g}_{m,n}^{\mathrm{B}} \right\|^2$. With diagonal approximation on the covariance matrix of NSGF $\{\widetilde{g}^{\mathrm{B}}\}$, i.e., $(\widetilde{g}_{m,n}^{\mathrm{B}})^H \cdot$

$\widetilde{g}^{\mathrm{B}}_{m',n'} = 0$ if $(m,n) \neq (m',n')$, we estimate the mask via

$$\min_{\boldsymbol{\sigma}^{\overrightarrow{\mathrm{AB}}}} \left\| C^{\mathrm{B}} - \boldsymbol{\sigma}^{\overrightarrow{\mathrm{AB}}} \odot C^{\mathrm{A}} \right\|^2 + \mu d(\boldsymbol{\sigma}^{\overrightarrow{\mathrm{AB}}}), \tag{13}$$

where $\odot$ denotes entrywise product. In this evaluation, we use the following regularization term

$$d(\boldsymbol{\sigma}^{\overrightarrow{\mathrm{AB}}}) = \left\| \boldsymbol{\sigma}^{\overrightarrow{\mathrm{AB}}} - \boldsymbol{\sigma}^{\mathrm{Ref}} \right\|_2^2 . \tag{14}$$

With (14), the objective function in (13) is a quadratic form of $\boldsymbol{\sigma}^{\overrightarrow{\mathrm{AB}}}$, which leads to the following explicit solution

$$\boldsymbol{\sigma}^{\overrightarrow{\mathrm{AB}}} = \frac{\overline{C^{\mathrm{A}}} \odot C^{\mathrm{B}} + \mu \boldsymbol{\sigma}^{\mathrm{Ref}}}{|C^{\mathrm{A}}|^2 + \mu}. \tag{15}$$

Here $\overline{}$ denotes complex conjugate.

5.1. Experimental settings

For experimental evaluation, we extracted two sets of English vowels, /iy/ and /u/[1], from the TIMIT database [22]. The vowels were from 390 speakers. For each speaker, there were 3 samples of /iy/ as well as 3 samples of /u/ included. The signals were down-sampled at 8000 Hz. Fundamental frequency was obtained with the method proposed in [20, 21] and assumed known throughout the evaluation.

We chose from the 390 speakers a reference speaker whose fundamental frequency was about the average of all speakers'. For the NSGT, we used Hann window with support interval of 20ms length. The time hop-size a was set to 4ms. For the pitch-dependent frequency hop-size, i.e., (8), we set $\mathsf{q} = 75$ according to pilot tests. For p, we used an average value of the first formant frequency ($F1$) as anchor frequency and the average $\bar{f}_0$ of a speaker as reference and fixed $\mathsf{p} = \lfloor F1/\bar{f}_0 \rceil$ for the speaker. We used $F1 = 280$ Hz and $F1 = 310$ Hz for /iy/ and /u/, respectively [23]. For (15), we empirically set $\boldsymbol{\sigma}^{\mathrm{Ref}} = \mathbf{1}$ (all-ones) and $\mu = 10^{-7}$. Part of the routines in the LTFAT toolbox [1, 24] were used to implement the NSGT.

For each vowel type, the frame masks for an input speaker were computed from 3×3 pairs of signals[2]. To obtain a variety of masks, for a signal pair we computed the frame masks as illustrated in Fig. 1. Hence, C^{A} and C^{B} in (15) were one-columnwise for the feature extraction. The obtained mask vectors were used as speaker feature vectors. We employed fully connected deep neural network (DNN) for the evaluation. The feature vectors were used as DNN input and they were divided in the following way for training and testing. For each speaker, $2/3$ of the speaker's masks were randomly selected as training data, and the rest $1/3$ were used for testing. The DNN structure was set as $1200-1024-1024-1024-390$ [3]. For DNN training, the following settings were used [25, 26, 27]. The number of epochs for RBM pre-training was 50, with learning rate set as 10^{-5}. The number of epochs for DNN fine-tuning was 25, where in the first 5 epochs only the parameters of the output layer were adjusted. The mini-batch size was set to 100.

5.2. Results

Fig. 2 shows performance of the harmonic-aligned frame mask (HAFM) in the vowel-dependent speaker classification tasks. For comparison, the (13-dimensional) mel-frequency cepstral coefficients (MFCC) [28] and the NSGT coefficients (C-NSGT) were also evaluated in the same way. We also tested the condition that f_0 was included as an extra feature dimension. It can be seen from the results that C-NSGT mostly performed the worst. On the other hand, HAFM which is established based on C-NSGT outperforms the others with noticeably higher accuracy. This implies that with the comparison way of feature extraction, the HAFM feature is more effective to capture and represent the speaker variations. The accuracy of HAFM is 83% for the "DNN/iy/+DNN/u/" case (i.e., DNNs of both vowels were combined for decision). It can also be noticed that to include f_0 as extra feature seems beneficial for MFCC. However, such benefit is generally not observed for both C-NGST and HAFM, as f_0 related information has already been well incorporated in these features.

In the evaluation, it was also observed that the frame mask based DNNs performed extremely well in distinguishing the reference speaker from the rest of the speakers. As the frame mask features were obtained by exhaustive comparison to the reference speaker, the resulted DNNs were inherently good verification models for the reference speaker.

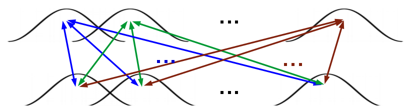

Figure 1: *Extraction of frame mask feature between two signals. (With only the time-shifted windows shown)*

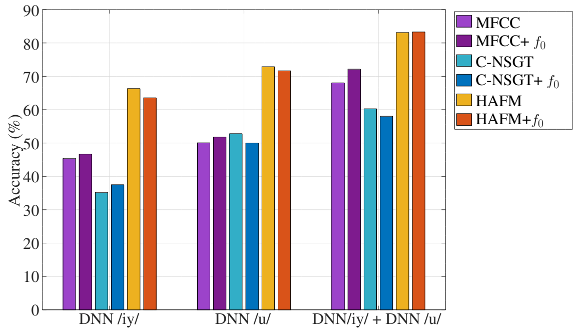

Figure 2: *Performance of DNN-based speaker classification. (The total number of DNN training data was roughly 1.5×10^5 for HAFM, and 2.5×10^4 for MFCC and C-NSGT.)*

6. Conclusions

The frame mask approach has been extended from instrumental sound analysis to voiced speech analysis. We have addressed the related issue by developing non-stationary Gabor transform (NSGT) with pitch-dependent and time-varying frequency resolution. The transform allows effective harmonic alignment in the transform domain. On this basis, harmonic-aligned frame mask has been proposed for voiced speech signals. We have used the proposed frame mask as similarity measure to compare and distinguish speaker identities, and have evaluated the proposal in a vowel-dependent and limited-data setting. Results confirm that the proposed frame mask is feasible for speech applications. It is effective in representing speaker characteristics in the content-dependent context and shows a potential for speaker identity related applications, specially for limited data scenarios.

7. Acknowledgements

This work was supported in part by the Austrian Science Fund (FWF) START-project (Y 551-N13).

[1] ARPAbet phonetic symbols were used in the database's documents. The symbols here may correspond to IPA symboles as /iy/–[i] and /u/(/ux/+/uw/)–[u].

[2] As there were also 3 samples from the reference speaker.

[3] Softmax layer was used as output layer, where an output node corresponded to a speaker ID. The input dimension was 1200.

8. References

[1] P. Søndergaard, B. Torrésani, and P. Balazs, "The linear time frequency analysis toolbox," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 10, no. 4, p. 1250032, 2012. [Online]. Available: http://ltfat.github.io/

[2] D. Gabor, "Theory of communication," *J. IEE - Part I: General*, vol. 94, no. 73, pp. 429–457, January 1947.

[3] S. Mallat, *A Wavelet Tour of Signal Processing - The Sparse Way*, 3rd ed. Academic Press, 2009.

[4] K. Gröchenig, *Foundations of Time-Frequency Analysis*. Boston, MA, USA, 2001.

[5] H. G. Feichtinger and K. Nowak, *A first survey of Gabor multipliers*, 2003, ch. 5, pp. 99–128.

[6] D. T. Stoeva and P. Balazs, "Invertibility of multipliers," *Applied and Computational Harmonic Analysis*, vol. 33, no. 2, pp. 292–299, 2012.

[7] P. Balazs and D. T. Stoeva, "Representation of the inverse of a multiplier," *Journal of Mathematical Analysis and Applications*, vol. 422, pp. 981–994, 2015.

[8] F. Hlawatsch, G. Matz, H. Kirchauer, and W. Kozek, "Time-frequency formulation, design, and implementation of time-varying optimal filters for signal estimation," *IEEE Trans. Signal Process.*, vol. 48, no. 5, pp. 1417 –1432, May 2000.

[9] P. Balazs, B. Laback, G. Eckel, and W. A. Deutsch, "Time-frequency sparsity by removing perceptually irrelevant components using a simple model of simultaneous masking," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 1, pp. 34–49, 2010.

[10] P. Majdak, P. Balazs, W. Kreuzer, and M. Dörfler, "A time-frequency method for increasing the signal-to-noise ratio in system identification with exponential sweeps," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP) 2011*, 2011.

[11] P. Depalle, R. Kronland-Martinet, and B. Torrésani, "Time-frequency multipliers for sound synthesis," in *Proc. SPIE, Wavelets XII*, 2007, pp. 221–224.

[12] A. Olivero, B. Torresani, and R. Kronland-Martinet, "A class of algorithms for time-frequency multiplier estimation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 8, pp. 1550–1559, Aug 2013.

[13] P. Balazs, M. Dörfler, F. Jaillet, N. Holighaus, and G. Velasco, "Theory, implementation and applications of nonstationary Gabor frames," *Journal of Computational and Applied Mathematics*, vol. 236, no. 6, pp. 1481 – 1496, 2011.

[14] N. Holighaus, M. Dörfler, G. A. Velasco, and T. Grill, "A framework for invertible, real-time constant-Q transforms," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 4, pp. 775–785, April 2013.

[15] E. S. Ottosen and M. Dörfler, "A phase vocoder based on non-stationary Gabor frames," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 11, pp. 2199–2208, Nov 2017.

[16] P. Casazza, "The art of frame theory," *Taiwanese J. Math.*, vol. 4, no. 2, pp. 129–202, 2000.

[17] O. Christensen, *An Introduction to Frames and Riesz Bases*. Birkhäuser Boston, 2003.

[18] H. G. Feichtinger and T. Strohmer, *Gabor Analysis and Algorithms - Theory and Applications*. Birkhäuser Boston, 1998.

[19] P. Balazs, "Basic definition and properties of Bessel multipliers," *Journal of Mathematical Analysis and Applications*, vol. 325, no. 1, pp. 571 – 585, 2007.

[20] F. Huang and T. Lee, "Pitch estimation in noisy speech using accumulated peak spectrum and sparse estimation technique," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 1, pp. 99–109, Jan. 2013.

[21] F. Huang and P. Balazs, "Dictionary learning for pitch estimation in speech signals," in *Proc. 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2017, pp. 1–6.

[22] "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993. [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html

[23] P. Ladefoged and K. Johnson, *A course in phonetics*, 6th ed. Boston, MA: Wadsworth, Cengage Learning, 2011.

[24] Z. Průša, P. L. Søndergaard, N. Holighaus, C. Wiesmeyr, and P. Balazs, "The Large Time-Frequency Analysis Toolbox 2.0," in *Sound, Music, and Motion*, ser. Lecture Notes in Computer Science, M. Aramaki, O. Derrien, R. Kronland-Martinet, and S. Ystad, Eds. Springer International Publishing, 2014, pp. 419–442.

[25] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[26] G. Hinton, "A practical guide to training restricted Boltzmann machines," Tech. Rep., 2010. [Online]. Available: http://www.cs.toronto.edu/ hinton/absps/guideTR.pdf

[27] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Letters*, vol. 21, no. 1, pp. 65–68, Jan 2014.

[28] Z. Fang, Z. Guoliang, and S. Zhanjiang, "Comparison of different implementations of MFCC," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, Nov. 2001.