Preliminary Study on SSCF-derived Polar Coordinate for ASR

Sotheara Leang, Eric Castelli, Dominique Vaufreydaz, Sethserey Sam

PRELIMINARY STUDY ON SSCF-DERIVED POLAR COORDINATE FOR ASR

AUTHOR VERSION

Sotheara Leang[1,2], **Eric Castelli**[2], **Dominique Vaufreydaz**[2], **Sethserey Sam**[1]

[1] Institute of Digital Research and Innovation, CADT, Phnom Penh, Cambodia
[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

ABSTRACT

The transition angles are defined to describe the vowel-to-vowel transitions in the acoustic space of the Spectral Subband Centroids, and the findings show that they are similar among speakers and speaking rates. In this paper, we propose to investigate the usage of polar coordinates in favor of angles to describe a speech signal by characterizing its acoustic trajectory and using them in Automatic Speech Recognition. According to the experimental results evaluated on the BRAF100 dataset, the polar coordinates achieved significantly higher accuracy than the angles in the mixed and cross-gender speech recognitions, demonstrating that these representations are superior at defining the acoustic trajectory of the speech signal. Furthermore, the accuracy was significantly improved when they were utilized with their first and second-order derivatives (Δ, $\Delta\Delta$), especially in cross-female recognition. However, the results showed they were not much more gender-independent than the conventional Mel-frequency Cepstral Coefficients (MFCCs).

***K*eywords** Automatic Speech Recognition, Spectral Subband Centroid Frequency, Speaker Normalization

1 Introduction

Automatic Speech Recognition (ASR) plays a vital role in human-computer interactions, particularly in voice assistants. It allows such intelligent devices to translate a speech signal into textual information to obtain semantic comprehension before taking action. Due to the rapid growth of related disciplines, ASR systems have remarkably improved and are extensively used in several sectors, significantly improving work efficiency and reducing human demands.

Most state-of-the-art ASR systems based on Hidden Markov Model (HMM) or Deep Learning characterize the speech signal with acoustic features derived from the absolute frequency measurements such as Mel-frequency Cepstral Coefficients (MFCCs), Perceptual Linear Predictive Cepstrum (PLP) and Mel-filter Bank [1, 2]. However, it is well known that the frequency space of speakers varies greatly, especially if we compare that of men with that of women or children: women and children produce speech with higher frequencies than men [3, 4, 5, 6, 7]. Furthermore, if the physiological, emotional, or sociological factors (*i.e.,* age, speaking rate, accent) are considered, the acoustic space varies significantly due to the inter-speaker and intra-speaker differences [8].

To take into account the large variability of speech (inter-speaker variability, environmental variability, etc.), current ASR systems require a very large amount of data for their training. This is even more true in the case of systems based on Deep Learning. This is a major handicap for developing such systems for poorly endowed or endangered languages. It is, in fact, often quite challenging and very costly to develop an ASR system for minority or low-resource languages when a large dataset of the transcribed speech does not exist [9, 10]. One research direction to reduce the training data size is to make the ASR systems intrinsically speaker-independent. If this is possible for the acoustic part of the systems, only one speaker would be sufficient for the learning phase.

This research proposes to describe the speech signal based on "acoustic gestures", with the hypothesis that it is more robust and capable of improving the performance of ASR systems, particularly with respect to the natural variability of the speech. We suggest defining the acoustic gestures of the speech signal using polar coordinates rather than the transition angles on the acoustic space of the Spectral Subband Centroid Frequency (SSCF). We anticipate that the polar coordinates provide a better representation of the speech because they can define not only the transition direction but also the acoustic trajectory of the speech signal.

This paper is structured as follows: Section 2 includes related work. The proposed method is described in Section 3. The experiments are introduced in Section 4. Section 5 contains the results and discussions. Finally, section 6 presents the conclusion and future work.

2 Related Work

The Spectral Subband Centroids (SSCs) were examined and utilized as complement parameters to the cepstral features for speech recognition. They have properties similar to the formant frequencies and are robust to additive Gaussian noise [11]. The SSC frequencies (SSCFs) were computed by dividing the frequency band into a number of subbands and computing the centroid of each subband using power spectrum of the speech signal (Equation 1).

$$SSCF_m = \frac{\int_{l_m}^{h_m} f w_m(f) P^\gamma(f)\, df}{\int_{l_m}^{h_m} w_m(f) P^\gamma(f)\, df} \quad (1)$$

Where l_m and h_m are the lower and upper bounds of subband m; w_m is the subband filter; $P(f)$ is the power spectrum at frequency bin f; γ is a coefficient regulating dynamic range of the power spectrum.

The SSCFs were subsequently explored in Vietnamese vowel-to-vowel (VV) transitions. Six subband filters were proposed to capture more information about the speech signal by splitting the frequency band in mel-scale into the equals-size subbands. The study found that the SSCF parameters have similar shape to the formant frequencies and provide a good estimation and continuous values, even during the consonant production [12].

The VV trajectories can be considered straight lines on the SSCF1-SSCF2 plane (which corresponds roughly to the F1-F2 plane) (see Figure 1). For each of these lines, it is then possible to define an angle with reference to the SSCF1 axis. For each pair of the transitions, V1V2 and V2V1, the sum of the two angles is equal to 180° (see Figure 2). Starting from this observation, the angles were proposed to characterize the direction of the vocalic transition between the adjacency frames on the SSCF planes. The angles were measured by hand over 80% of the transition duration. According to the study, the angles were asserted to be speaker-independent because they were roughly the same across diverse speakers, including men and women, and similar results were obtained for the different speaking rates (see Figure 3) [12].

3 Proposed Method

Previous work assumed that the spectral trajectories of vowel-to-vowel transitions are quasi-straight lines on SSCF1-SSCF2 plane. Thus it is possible to compute the angles to define the vocalic transitions [12]. The angles are generalized to calculate across all planes of $SSCF_i$-$SSCF_{i+1}$.

In each $SSCF_i$-$SSCF_{i+1}$ plane, the angle between a transition window of N frames are defined as in Equation 2.

$$Angle(j) = \frac{\pi}{180^\circ} \arctan\left(\frac{\Delta SSCF_{i+1}}{\Delta SSCF_i}\right) \quad (2)$$

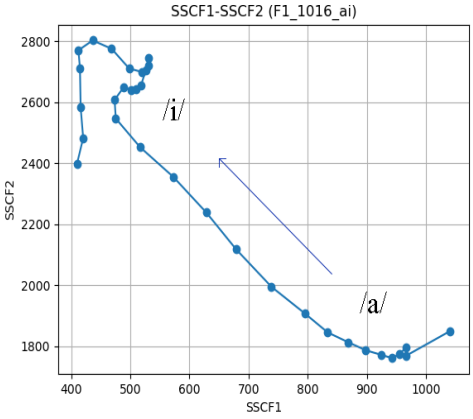

Figure 1: The spectral transition of /ai/ on the SSCF1-SSCF2 plane produced by a Vietnamese female speaker

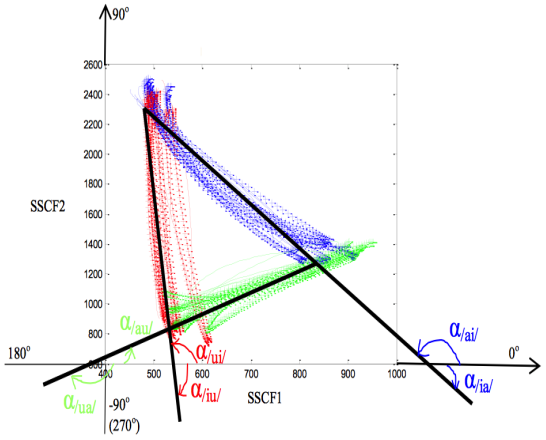

Figure 2: The SSCF angles of six vowel-to-vowel transitions on the SSCF1-SSCF2 plane produced by a Vietnamese male speaker (source [12]).

Where j is a frame at index j; $\Delta SSCF_i$ is the difference between SSCF on the ith axis at the end and the beginning of the transition.

The transition angles are computed directly from the arctan function, which has a discontinuity at $-\pi$ and $+\pi$, resulting in the steps throughout the transition (see Figure 4). Therefore, obtaining the angle directly from the arctan function is not a continuous parameter. Our experiments showed that using only arctan angles to define the acoustic trajectory of the speech signal is probably not a very good idea because the jumps will produce a kind of "noise" in the data.

That is why we propose to characterize the spectral trajectory in the acoustic space of the SSCFs using the polar coordinates. We assert that they give a more accurate representation because they are continuous variables and can specify not only the transition direction but also the tra-

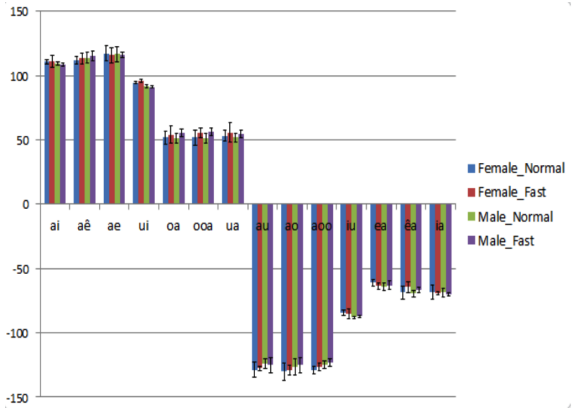

Figure 3: The SSCF angles on the SSCF1-SSCF2 plane at fast and normal rates produced by a Vietnamese male and female (source [12]).

jectory of the spectrum. For a given frame, the two polar parameters (angle and radius) on the $SSCFi\text{-}SSCF_{i+1}$ plane are defined as in Equation 3.

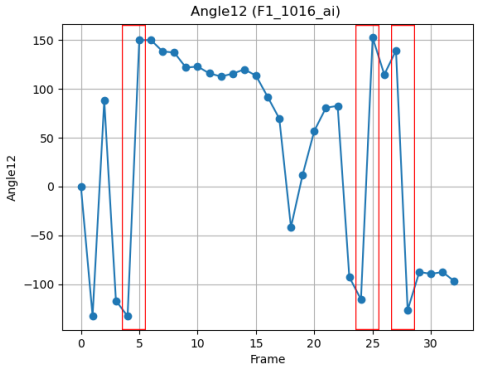

Figure 4: The transition angles of /ai/ on the SSCF1-SSCF2 plane produced by a Vietnamese female speaker. The steps caused by the arctan function are highlighted in red rectangular.

$$
\begin{aligned}
Angle(j) &= \frac{180^\circ}{\pi} \arctan\left(\frac{SSCF_{i+1}}{SSCF_i}\right) \\
Radius(j) &= \sqrt{SSCF_{i+1}^2 + SSCF_i^2}
\end{aligned}
\tag{3}
$$

Where j represents a frame at index j; $SSCF_i$ and $SSCF_{i+1}$ represent the axis of the SSCF plane, respectively.

4 Experiments

The study made use of the Kaldi toolkit [13]. Note that the language model was trained using SRILM [14] on the transcripts of the speech corpus because we want to measure the effect of the acoustic model, not the effect of out-of-vocabulary (OOV) words that existed in the test sets. The lexicon was constructed from the vocabulary of the language model using Phonetisaurus [15].

Table 1: Specification of BRAF100

Corpus	Hour	Transcript	Speaker	Vocab
BRAF100	28h	10,564	100	22,199

4.1 Datasets

The BRAF100 corpus [16] was used for French speech recognition. It contains around 28 hours of recordings from 100 native speakers (50 are men) aged from 15 to 63. Each speaker recorded between 102 and 105 different sentences and a common extract of "La Science et l'Hypothèse" of Henry Poincarée. The specification of the corpus is given in Table 1.

The corpus was prepared into three sets, each containing the train and test sets with comparable speaker sizes. The first set is for normal speech recognition, and the ASR model was trained and evaluated on the dataset of male and female speakers. The remaining two sets are for cross-gender speech recognition. The ASR model is trained on a male dataset and evaluated on a female dataset in cross-male recognition, whereas cross-female recognition does the opposite. The specification of datasets are specified in the Table 2.

4.2 DNN-HMM Model

The ASR systems were constructed using a hybrid DNN-HMM model. The overall architecture of the DNN is given in Figure 5. The network takes nine consecutive contextual frames as the input and feeds them into three hidden layers, each consisting of 512 neurons, and the hyperbolic tangent function (tanh) was employed as an activation function. These hidden layers are used to learn the representation of the spliced input features. The output from the hidden layers was then fetched into a fully connected (FC) layer to estimate the state of the triphone HMM using the softmax function.

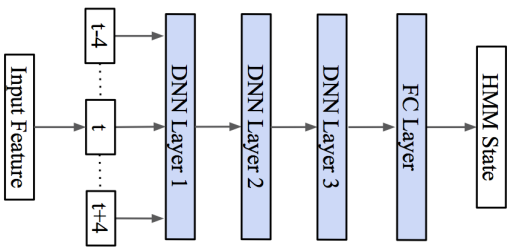

Figure 5: Overview architecture of the DNN

Table 2: The experimental datasets constructed from BRAF100

Experiment	Train Set			Test Set		
	Hour	**Transcript**	**Speaker**	**Hour**	**Transcript**	**Speaker**
Normal recognition	12h	4,657	22 males, 22 females	3h	1,036	5 males, 5 females
Cross-male recognition	12h	4,661	44 males	1.5h	519	5 females
Cross-female recognition	13h	5,278	44 females	1.5h	517	5 males

Three training steps were conducted to produce the DNN-HMM model. First, a context-independent model (Monophone) was trained on the input features using 1K Gaussians over 40 iterations. Then, context-dependent (Triphone) training was performed over 35 iterations, utilizing 2K states and 10K Gaussians. Finally, the triphone DNN training was conducted across 20 epochs on the aligned training data from the context-dependent model. The initial and final learning rates of 0.01 and 0.001 were used during the training.

4.3 Feature Configuration

In this study, the angles and the polar coordinates were evaluated for characterizing the spectral trajectory of the speech signal on the SSCF planes and compared to the classical MFCC, the most widely used acoustic feature for speech recognition.

The angles were defined over a window size of one. The SSCFs were computed from short-time frames of 25 ms length with a 10 ms overlap, and the moving average was applied across a three-frame window to reduce the influence of the noise. A pre-emphasis factor of 0.97 and a Hamming window were utilized during feature extraction.

The MFCC with 6 and 13 dimensions were used as the baseline features. They were computed using 6 and 13 subband filters on the short-time frames with the same configuration as SSCFs, and a cepstral liftering coefficient of 22 was used while extracting the features.

5 Results and Discussions

The results in word error rate (WER) of normal and cross-gender speech recognitions are presented in Table 3. In normal recognition cases, the MFCC significantly outperformed the angle and polar coordinate. As expected, the 13 MFCC produced a greater result than the 6 MFCC because it captured more signal information (higher precision in the frequency domain). In addition, the performance was marginally improved when speed and acceleration were used ($\Delta, \Delta\Delta$).

The angle had the greatest error rate of 94%, demonstrating that it is not a good parameter due to its discontinuity. However, the estimation of transition trajectory using polar coordinate surpassed angle by a wide margin. Furthermore, the performance was significantly improved when it was combined with speed and acceleration, particularly in cross-female recognition.

In the transgender recognition cases, similar patterns were obtained if the different types of parameters were compared. Nevertheless, all parameters had significantly higher error rates in all cases. It can be seen that the recognition rates for one gender are poorer if the learning is performed with speakers of the other gender. This shows that these parameters are not really independent of gender.

In cross-male recognition, the 13 MFCC performed worse than the 6 MFCC, and the situation was quite similar when used with delta features ($\Delta, \Delta\Delta$). However, this was not the case for cross-female recognition.

Overall, we still find that recognition rates are similar between the polar parameters and the MFCC (particularly for the 6 MFCC) compared to the angles.

However, the polar parameters characterizing the spectral trajectories do not seem much more gender-independent than the classical MFCC. This may be due to several aspects that we have not yet investigated in this first study:

1) We considered SSCF0 (roughly corresponding to the fundamental frequency F0) participating in the same way as the other SSCFi similar to the formants. However, it is well known that the fundamental frequency F0 is quite different between men and women. Moreover, calculating a trajectory angle between SSCF0 and SSCF1 (F0 and F1) is probably not judicious because it does not correspond to real physical phenomena that SSCF0 (F0) characterizes the vibration of vocal cords, whereas the other SSCF characterize the resonances of the vocal tract.

2) A second point to investigate concerns the fact that characterizing the trajectories only by angles. If this is valid in the SSCF1-SSCF2 (F1-F2) plane where the trajectory during vowel-vowel transitions is indeed a quasi-straight line, it is not the case for the other $SSCFi$-$SSCF_{i+1}$ planes for $i > 2$. We must consider characterizing this trajectory by another means.

3) A dynamic gesture is classically characterized by its trajectory and speed (and acceleration) of the movement on the trajectory. It is therefore important to characterize velocity and acceleration more judiciously than simply computing the first and second derivatives ($\Delta, \Delta\Delta$).

Table 3: Experimental results on BRAF100 in WER(%)

Experiment	Parameter	Dimension	MONO	TRI	DNN
Normal recognition	6 MFCC	6	62.00	31.24	10.90
	6 MFCC, Δ, ΔΔ	18	26.70	11.20	08.65
	13 MFCC	13	41.90	17.09	08.39
	13 MFCC, Δ, ΔΔ	39	20.83	09.10	07.14
	Angle	5	97.12	94.02	94.19
	Polar	10	71.62	42.35	16.44
	Polar, Δ, ΔΔ	30	46.17	21.10	13.01
Cross-male recognition	6 MFCC	6	67.41	39.58	16.51
	6 MFCC, Δ, ΔΔ	18	32.19	14.84	11.51
	13 MFCC	13	62.65	41.04	20.58
	13 MFCC, Δ, ΔΔ	39	36.39	23.12	16.74
	Angle	5	96.87	97.29	94.60
	Polar	10	79.25	56.81	28.43
	Polar, Δ, ΔΔ	30	57.85	33.86	21.51
Cross-female recognition	6 MFCC	6	73.71	51.54	20.44
	6 MFCC, Δ, ΔΔ	18	35.63	17.59	15.30
	13 MFCC	13	66.96	43.92	18.62
	13 MFCC, Δ, ΔΔ	39	32.55	16.49	13.45
	Angle	5	93.78	94.15	93.51
	Polar	10	85.73	74.24	41.09
	Polar, Δ, ΔΔ	30	64.31	48.24	29.20

6 Conclusion

In this work, we investigated the use of polar coordinates in SSCF planes to describe the speech signal based on its acoustic trajectory. The finding showed that they significantly outperformed angles in both normal and cross-gender recognitions, demonstrating that the polar representation provides better information in characterizing the acoustic elements of the speech signal.

Even though the polar coordinates produced more significant results, there is still room for improvement. Our experiments showed that they are not significantly more gender-independent than the classical MFCC. We propose several directions for investigation in order to better account for the dynamic aspects of speech production. Examining how the cochlear system works will provide us with more useful ideas for characterizing the time aspect. Furthermore, it is desirable to investigate how to efficiently manage SSCF0 because using it directly with higher SSCFs is irrelevant. Finally, the SSCF0 must be normalized to account for spectral variation between speakers, particularly between male and female voices.

Acknowledgment

This work was carried out as part of a PhD thesis funded by the French Embassy in Cambodia and the Cambodia Academy of Digital Technology.

References

[1] Suman K Saksamudre, PP Shrishrimal, and RR Deshmukh, "A review on different approaches for speech recognition system," *International Journal of Computer Applications*, vol. 115, no. 22, 2015.

[2] Divya Gupta, Poonam Bansal, and Kavita Choudhary, "The state of the art of feature extraction techniques in speech recognition," *Speech and language processing for human-machine communications*, pp. 195–207, 2018.

[3] Pooja V Janse, Smita B Magre, Pratik K Kurzekar, and R Deshmukh, "A comparative study between mfcc and dwt feature extraction technique," *Inter-*

national Journal of Engineering Research and Technology, vol. 3, no. 1, pp. 3124–3127, 2014.

[4] Lucie Ménard, Jean-Luc Schwartz, Louis-Jean Boë, Sonia Kandel, and Nathalie Vallée, "Auditory normalization of french vowels synthesized by an articulatory model simulating growth from birth to adulthood," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1892–1905, 2002.

[5] Alexandros Potamianos and Shrikanth Narayanan, "Robust recognition of children's speech," *IEEE Transactions on speech and audio processing*, vol. 11, no. 6, pp. 603–616, 2003.

[6] James M Hillenbrand and Michael J Clark, "The role of f 0 and formant frequencies in distinguishing the voices of men and women," *Attention, Perception, & Psychophysics*, vol. 71, no. 5, pp. 1150–1166, 2009.

[7] René Carré, Pierre Divenyi, and Mohamad Mrayati, "Speech: A dynamic process," in *Speech: A dynamic process*. de Gruyter, 2017.

[8] Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al., "Automatic speech recognition and speech variability: A review," *Speech communication*, vol. 49, no. 10-11, pp. 763–786, 2007.

[9] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech communication*, vol. 56, pp. 85–100, 2014.

[10] Chongchong Yu, Meng Kang, Yunbing Chen, Jiajia Wu, and Xia Zhao, "Acoustic modeling based on deep learning for low-resource speech recognition: An overview," *IEEE Access*, vol. 8, pp. 163829–163843, 2020.

[11] Kuldip K Paliwal, "Spectral subband centroid features for speech recognition," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*. IEEE, 1998, vol. 2, pp. 617–620.

[12] Thi-Anh-Xuan Tran et al., *Acoustic gesture modeling. Application to a Vietnamese speech recognition system*, Ph.D. thesis, Université Grenoble Alpes (ComUE), 2016.

[13] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[14] "The sri language modeling toolkit," `http://www.speech.sri.com/projects/srilm`.

[15] "Phonetisaurus g2p," `https://github.com/AdolfVonKleist/Phonetisaurus`.

[16] Dominique Vaufreydaz, Carole Bergamini, Jean-François Serignat, Laurent Besacier, and Mohamad Akbar, "A new methodology for speech corpora definition from internet documents," in *LREC'2000 (Language Resources & Evaluation international Conference)*, 2000, pp. pp–423.