PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Madurese Speech Synthesis using HMM

Roudhotul ROUF; Dhany ARIFIANTO
Institut Teknologi Sepuluh Nopember, Indonesia

ABSTRACT

This research is intended to study Madura language which may be the only local language in Indonesia classified into a tonal language. However, the Madurese is not only under- documented in term of phonetics but under-resourced as well. The first step was developing the Madura language voice database. The initial results are limited to the fundamental frequency contour of native male and female utterances. We used a well-known technique called Mel Frequency Cepstral Coefficient (MFCC) to obtain acoustic cues, and the cues were further processed by observing the delta, Δ Cepstrum for velocity change in an utterance and the delta-delta Δ2, to indicate the acceleration or deceleration of the acoustical cues change over time, respectively. We used instantaneous frequency tracking in STRAIGHT. We obtained results were compared to the well-established TIMIT database in English and Indonesian Speech Database. From the initial results, the Δ and Δ2 indicated that the Madurese showed a rapid change in both time- and frequency- domain cues. Although the findings may be far from conclusive because the Madura island has four different regions that have its own accent that slightly different from each other. Currently, the on-going research aim is towards a high-quality Madurese speech synthesis.

Keywords: Madurese, Fundamental Frequency, STRAIGHT

1. INTRODUCTION

In all parts of the world, members of ethnolinguistic minorities are increasingly abandoning their native language in favour of another language, including in childrearing and formal education. Among ethnolinguistic communities, a variety of opinions on the future prospects of their languages can be observed. Some speakers of endangered languages come to consider their own language backward and impractical. Such negative views are often directly related to the socioeconomic pressure of a dominant speech community. Other speakers of endangered languages, however, attempt to directly counter these threats to their language, and commit themselves to language stabilization and revitalization activities. So according to the UNESCO 2003 this is being an international problem.

Indonesia is one of the countries that has a very high cultural diversity, differences in regional languages such as the Madura language are already thinning. One of the factors causing the extinction is due to the process of globalization and urbanization which can lead to the assimilation and acculturation of culture, especially in urban areas. So it needs a solution to maintain Madura Language, Many Madurese people are lacking in the level of concern for the diversity of Madura languages, public awareness that Madura language has its own unique level like Madura phonemes with pronunciation of exhaled words such as: bh, dh, gh, and jh [1]. Madurese is one of the local languages of Indonesia that's have been classified as under-documented language in term of phonetics and under resourced as well. As the age of Madura language continued to grow, the native speakers of the language were diminishing. If this continues, then not only Madura language, but even other regional languages that are the original wealth of the country will disappear because it becomes a dead language. In order to save so that the variety of regional languages did not become extinct. This research is intended to study Madura language

2. Madura Language Characterization

2.1 Vocals

According to KBBI, vocal sound is the sound of language produced by the flow of air from the lungs through the vocal cords and constriction of the sound channel above the glottis. Madura language has a special character in order to be easily read by Madurese or not Madurese [2].

Tabel 1 – *vocals phonems*

vowel	**Example of use in the madura language**		
	at the beginning	in the middle	at the end
a	*a*los 'smooth'	p*asa*r 'market'	Sal*a* 'false'
â	*â*poy 'fire'	ab*â*s 'see'	Bâb*â* 'under'
e	*e*ppa 'father'	neser 'pity'	-
è	*è*ntar 'go'	s*èsè*k 'slice'	Tal*è* 'rope'
i	*i*yâ 'yes'	raddhin 'beautiful'	Mandh*i* 'efficacious'
o	*o*lok 'call'	dokar 'gig'	Pa*o* 'mango'
u	-	dhuri 'thorns'	Pak*u* 'nail'

2.2 Consonant

According to the articulation, consonants in Madura language can be categorized based on four factors, namely: (1) the state of the vocal cords (2) the area of articulation (3) the way of articulation and (4) the presence or absence of aspirations. Based on the state of the vocal cords, consonants are divided into voiceless and voiceless consonants. Based on the area of articulation, consonants are distinguished by bilabial, labiodental, alveolar, palatal, velar and glottal consonants. Based on the method of articulation, the consonants are distinguished from inhibitory, fricative, nasal, vibrating, and lateral consonants. Based on the presence or absence of consonant aspirations distinguished from conspiracy aspirations and not aspirations. Furthermore, there are more semi-vocal forms, namely the sound of language which practically includes consonants, but seen from the articulation it has not formed a pure consonant (3). Madurese has 20 consonants, namely: *b, c, d, f, g, h, j, k, l, m, n, p, r, s, t, v, w, x, y, z.*

2.3 Combined Phonems

In the Madurese, the alphabets which represent phonemes are shown in the Table 2 as follows:

Table 2 – *Combined Phonemes*

vowel	**Example of use in the Madurese**		
	at the beginning	in the middle	at the end
kh	khusus 'special'	akher 'final'	tarekh 'provisions of time'
ng	ngelloh 'complain'	bunguh 'purple'	-
ny	nyatah 'real'	tanyo' 'washed away'	-
sy	syarat 'term'	isyarat 'sign'	-
bh	bhârâ 'lungs'	cabbhi 'chili'	-
dh	dhara 'dove'	pendheng	-
gh	ghibeh 'bring'	bighi 'seed'	-
jh	jhârâ 'horse'	bâjhâ 'steel'	-

2.4 Madurese Diphthongs

Diphthong is a sound formed by the combination of two vowels in a single syllable, in which the sound begins as one vowel and moves toward another. The Madurese has 4 diphthongs as follows:

Tabel 3 – Madurese Diphthong

ay	tapay 'tapay'	oy	tamoy 'guest'
ây	ghebây 'make'	uy	kerbhuy 'buffalo'

2.5 Fundamental Frequency

The sound formation process starts from the vibrating vocal cords (vocal cord and vocal fold) in the larynx due to the flow of air passing. The air flow is cut by the movement of the vocal cords into a pulse signal that is quasi-periodic, resulting in a vibrating frequency called the fundamental frequency (fundamental). This typical frequency is influenced by the physiological conditions of the human larynx. Under normal conditions of conversation, habitual pitch levels range from 50Hz to 250Hz for men and 120Hz to 500Hz for women [4]. This frequency changes constantly and gives someone linguistic information such as the difference between intonation and emotion [5].

The fundamental frequency has an excitation parameter which is sound state information that aims to distinguish between sound regions (produce acoustic parameters) or not sound (pause). Acoustic sound parameters are obtained from spectral parameters which are information on physical quantities of sound that refer to mel-cepstrum from database sounds such as MFCC, delta cepstrum, delta-delta cepstrum and duration [6].

2.6 STRAIGHT

STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of Weighted Spectrum**),** This method is one method of analysis of extraction of F_0 values automatically [7]. This method manipulates adaptive spectral sound signals that are refined and extraction of F_0 based on fixed-point instantaneous-frequency. Fixed-points are based on the time-warping STFT method (short-term Fourier Transform-based) that deals with harmonic components to estimate minimum errors. In this method there are 2 stages, namely:

1. Band pass filter is performed which is proportional to the log frequency axis used. This stage is used to extract the fix point mapping from the middle frequency filter as the output filter.
2. The development process, as an estimate on the axis of F_0 and derivative F_0. In this stage extraction is done using F_0 using the STFT method.

The algorithm in extraction F_0 uses the instant frequency of non-stationary time i.e., Hilbert Transform H $[x(t_0)]$, for the filter design ω_s (t, λ) designed from Gabor function ω (t, λ). Where both of these functions are carried out by convolution with the order of the base function B-spline 2 (t, λ). The F_0 estimate is chosen from the fixed-point in the form of the maximum carrier to noise ratio (C / N ratio) (20dB or higher as the basic component). With mathematical models as follows:

$$\omega_s(t,\lambda) = \omega(t,\lambda) * \text{ɧ}(t,\lambda) \quad (1)$$

$$\omega(t,\lambda) = e^{-\frac{\lambda^2 t^2}{4\pi\eta^2}} e^{j\lambda t} \quad (2)$$

$$\text{ɧ}(t,\lambda) = max\left\{0{,}1 - \left|\frac{\lambda t}{2\pi\eta}\right|\right\} \quad (3)$$

where * is a symbol of convolution, η is a time elongation factor, F_0 is Basic Frequency

The basic concept of STRAIGHT is F_0 Extraction, aperiodicity, refined spectral envelope extraction. Aperiodicity expresses a measure of harmonious information comparisons of non-harmonic information in the frequency domain and can be expressed as the relative distribution of energy from non-harmonious components. This STRAIGHT method has high flexibility in manipulating speech signals regardless of the color tone, while maintaining high quality. In addition, this method for investigating quasi-periodic (8).

2.7 Cepstrum Distance

Cepstrum distance is a measurement of Euclidian distance from the log-spectral coefficient

between two frames. Cepstral coefficients can also be obtained from the results of linear prediction coefficients (LPC) [9]. The smaller value of the distance indicates that the two sounds are similar. The cepstrum distance equation is as follows:

$$d_c^2(Q) = \sum_{n=1}^{Q} (c_n - c_n')^2 \quad (4)$$

where d (Q) is distance, c_n is the first spectral signal value and c_n' is the second signal spectral value.

2.8 Experiment

The pre-preparation section, Data collection begins with the search for utterance, who has great qualifications and lives in Madurese from birth to the minimum of high school graduation. The utterance in the research amounted to 1 male (IIW) and 1 female (RJR). Then the data retrieval process is done by recording the voice of utterance.

2.8.1 Mechanism of Data Collection

Record voice of Madurese speakers using DBXRTA-M (dBx RTA Measurement Microphone) and USB Audio Interface in Focusrite. The microphone is placed 10-15 cm from the mouth of the speaker, so that there is no clipping signal. Clipping signal is a condition where the captured sound signal overlaps the amplitude in the area of sound signal analysis, this condition occurs when the recording device is too close to the sound source and the sound source has high energy so that if this data occurs the record cannot be used because it will occur losing some important information when analyzing sound signals. Recording is conditioned in the sampling frequency of 44100Hz and 24 bits rate. The things that must be considered when the preconditions of the recordings are obtained, namely: background noise and reverberation that are not large, there is no data clipping, cutting per voice segment, downsampling was carried out on the crying sound signal to 8000Hz.

2.8.2 Extraction Process

After recording, the recorded data will be continued with the process of segmentation and labeling using audacity software. The results of segmentation and labeling will be the material to be processed to obtain results in the form of basic frequencies and sound acoustic parameters (mel-cepstrum from database sounds such as delta cepstrum, delta-delta cepstrum).

Fundamental frequency extraction is done by the STRAIGHT method. The process of processing this method aims to obtain the results of estimating the fundamental frequency for a Madurese signal that has different characteristics from the comparative language (Indonesian and English). The estimation of the cepstral and double delta cepstral values is obtained from the cepstrum estimation value of the MFCC method. The order value used is the value taken from the order standard for sound synthesis.

2.8.3 F0 Data Analysis Fundamental Frequency

The results of F_0 extraction in the form of a spectrogram which is a reference in the F0 estimation in Madurese speech sound signals are shown in Figure 2.

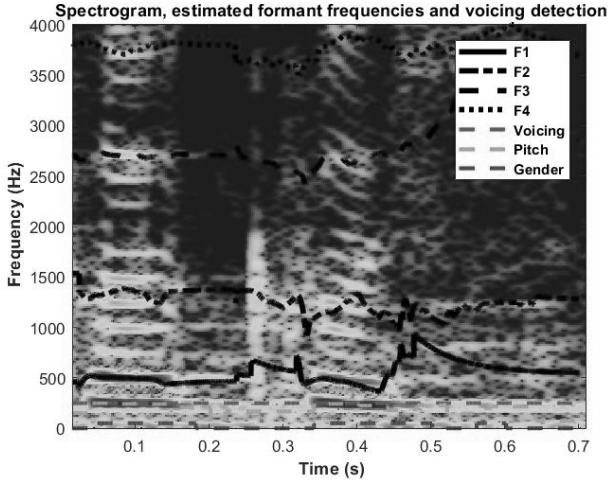
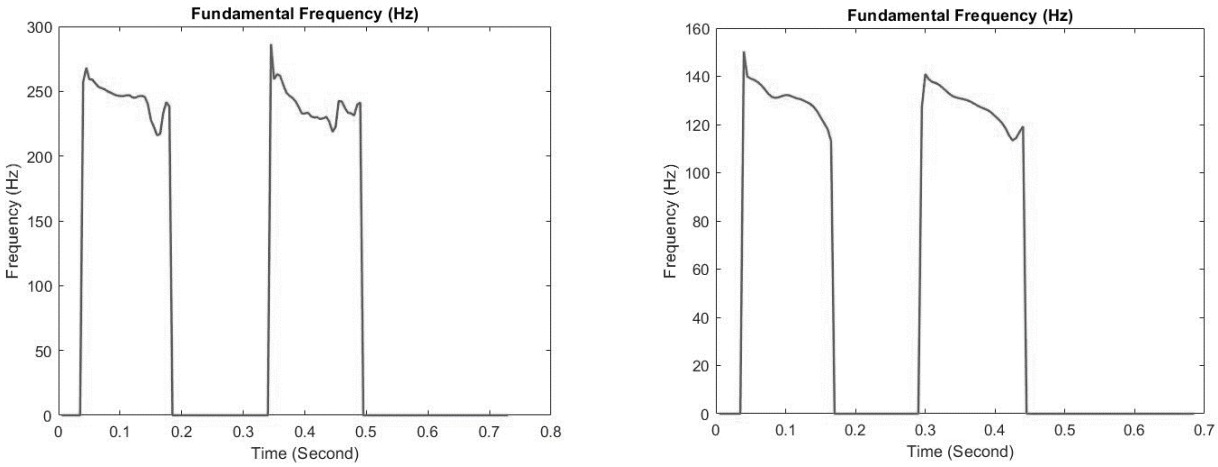


Figure 2 - Histogram Plot of Male Utterance (RJR) " Bhâghus"

Besides that, the results of F_0 extraction in the form of a pinch which is a reference in F_0 estimation in some samples of words and phonemes of both male and female sound signals are shown in Fig. 3.



(a) (b)

Figure 3 - F_0 *plot sample of word " Bhâghus" (a) female (RJR) and (b) male (IIW)*

2.8.4 Cepstrum Distance Analysis

Cepstrum Distance Results from the comparison 2 utterances between Madurese speakers and comparative language (Indonesian and English) are produced as follows:

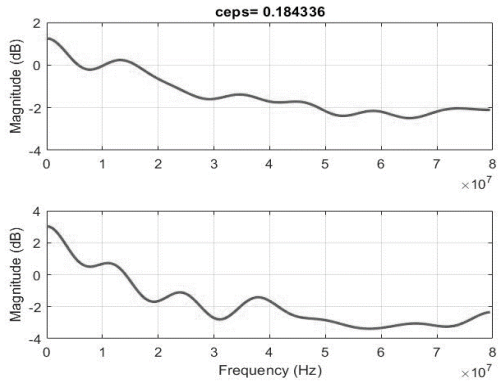


Figure 4. *Cepstrum Distance Madura language of Female and Male Utterance*

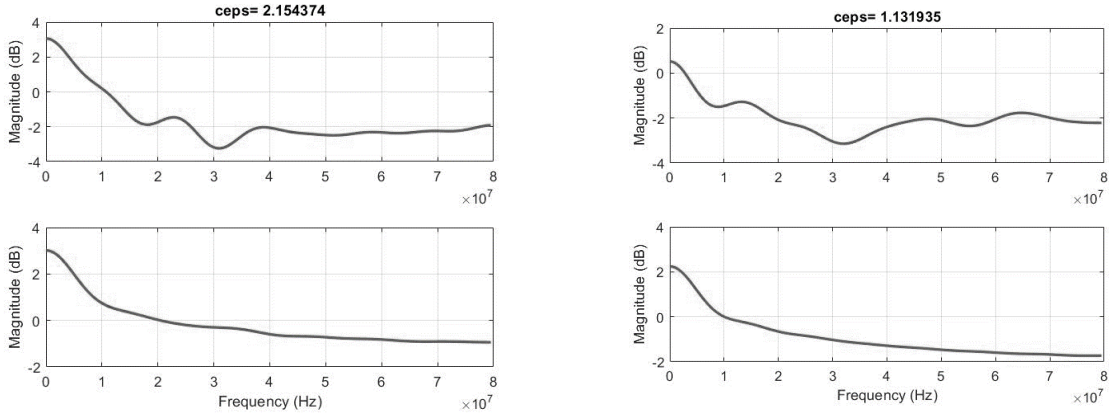
Cepstrum distance is the result of the comparison of two utterance above, showing the Cepstrum distance <1 that the two fundamental frequencies (F0) of the two speakers have similarities.

Tabel 3 - Cepstrum Distance Madura language of Female and Male Utterance

No.	Female Utterance	Male Utterance	CEPS
1.	tanəm	tanəm	0.262026
2.	dhəbu	dhəbu	0.000281
3.	Dhəmar	Dhəmar	0.121080
4.	ghəmpang	ghəmpang	0.001179
5.	settɔŋ	settɔŋ	0.336879

The results of the comparison of the two records in the above word sample show a Cepstrum

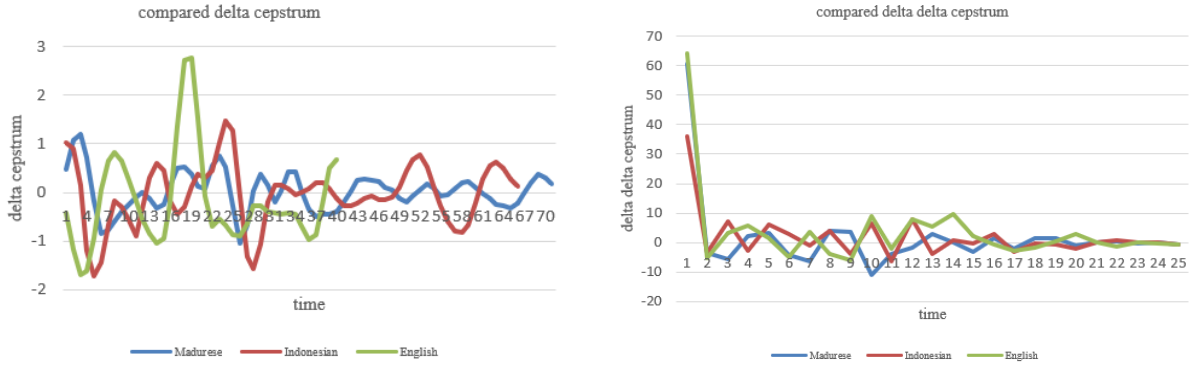
distance <0.3, which indicates that the two fundamental frequencies (F0) of the two speakers have similarities.
The following are the results of a comparison of the basic frequency and results of the distance between Madurese, Indonesian and English by female and male speakers, which is shown in the Fig. 4 below.



(a) (b)

Figure 4 - *Estimated results with cepstrum distance on comparative (a) Madurese - Indonesia, (b) Madurese – English*

2.8.5 Prosody Parameter Data Analysis



(a) (b)

Figure 5 - *Estimated results with (a) delta cepstrum and (b) delta-delta cepstrum of maduranese utterance "bhəghus" (blue), Indonesian utterance "berusahalah" (red) and English Utterance"the straw" by KI (green)*

3. Discussion

Fundamental frequency analysis has been done to get the characteristics of Madura Language utterance which its sound signal, such as the average fundamental frequency of adult human voice based on gender, such as for male (50Hz - 200Hz) lower than the female fundamental frequency (150Hz - 400Hz). Based on the fundamental frequency estimation results shown in Figure 3 shows that the Madura language narrative has the same pinch pattern, the results shown also obtained large fundamental frequencies of Madura language speakers, namely the fundamental frequency of RJR (Female Utterance) and IIW (Male Utterance) are in the range of 150-300 Hz and 50-200 Hz according to the standard. which is on the basis of the theory where the fundamental frequency of women is higher than that of men. In Table 3 show that cepstrum distance Madurese of female (RJR) and male (IIW) between Madurese utterance < 0.3, which indicates that the two fundamental frequencies (F0) of the two speakers have similarity. Whereas when Madura language compared with other languages have the Cepstrum Distance > 1 which states that the fundamental frequency (F0) of the two speakers voices, both between Madurese and Indonesian or English, have no resemblance. This happened

between female and male speakers as seen in Figure 4.

Analysis of prosody parameters for this language based on its dynamic features of a signal analysis. In this analysis, the MFCC (Mel Frequency Cepstral Coefficient) method is used to estimate the cepstral coefficient, the value of delta cepstrum (ΔF0) (peak difference from cepstrum) as information on the speed of change of sound and double delta value (Δ2F0) which is a derivative of F0 as acceleration the sound signal frequency obtained from the acceleration of the displacement between cepstrum peaks. In Figure 5.a, the tone of ΔF0 value indicates the speed of change from cepstral value for each Language type. From the three languages that have been compared above, it can be concluded that the value of ΔF0 which indicates of the speed for the utterance of each language sorted from the fastest is English, Madura and Indonesia. In his narrative, it can also be distinguished subjectively that Madura language is faster than Indonesian but slower than English. Whereas the value of Δ2F0 is the result of estimating changes in frequency acceleration in each Language type shown in Figure 5.b as shown in the analysis of the value of ΔF0 that the change in frequency acceleration for each Language type above has a tendency to not change the acceleration between Languages. The value of Δ2F0 indicates the speakers' intonation. The value of Δ2F0 from the Madurese should appearing his accent and have an acceleration that is different from the other languages. However, when sampling is done on every word that tends not to give accent from Madurese, it seems like speaking words in Madurese in a flat (without expressive) language. Therefore, the results of comparisons between the three languages above show that there is almost no change in intonation.

4. CONCLUSION

Based on the fundamental frequency estimation results show that the Madura Language narrative has the same pinch pattern, the magnitude of the fundamental frequency of RJR (female utterance) and IIW (male utterance) are in the range 150-300 Hz and 50-200 Hz, respectively. The Cepstrum Distance results show a Ceptrum distance between Madura languages < 0.3, which indicates that the F_0 of the two Madurese utterance have similarities. Whereas when compared between Madura language and other languages have the results of Cepstrum Distance > 1 which states that the F_0 of the two speakers' voices, both between Madurese and Indonesian or English, have no resemblance. Of the three languages that have been compared above, it can be concluded that the value of ΔF_0 which indicates the speed in the narrative of each of the fastest languages in a row is English, Madura and Indonesia. The analysis of $\Delta^2 F_0$ value indicates that the value of the change in frequency acceleration for each type of language above has a tendency that there is no change in acceleration (almost no change in intonation) between the languages.

ACKNOWLEDGEMENTS

Authors would like to thank for speakers as Madurese utterance for this research, wins recording studio who took a part in recording process of Madurese speech database. Thanks for Madurese expert of Airlangga University who reviewing Madurese speech database.

REFERENCES

1. Qutsiyah, F. H. Rachman and F. Solihin, "APLIKASI TEXT TO SPEECH DALAM SISTEM PENERJEMAH BAHASA INDONESIA-MADURA MENGGUNAKAN METODE FSA (FINITE STATE AUTOMATA)," in *JURNAL SARJANA TEKNIK INFORMATIKA*, Bangkalan, 2015
2. Rahilah, "Aplikasi Penerjemah Bahasa Madura-Indonesia dan Indonesia-Madura," in *Skripsi Jurusan Teknik Informatika Fakultas Teknik Universitas Trunojoyo Madura*, Bangkalan, 2013.
3. Sofyan, A., “Tata Bahasa Madura,” Sidoarjo: Balai Bahasa Surabaya. Sidoarjo: Balai Bahasa Surabaya, 2008.
4. Utomo, A. B., Wahyudi, dan Hidayanto, A., “Analisa Karakteristik Suara Manusia Berdasarkan Frekuensi Fundamental, dan Tingkat usia pada pelajar SLTP dan SMA,” Makalah Tugas Akhir Jurusan Teknik Elektro, Universitas Diponegoro, 2007.
5. Al-Azhar, M. N., 2011. *Audio Forensic : Theory And Analysis*. Pusat Laboratorium Forensik Polri Bidang Fisika Dan Komputer Forensik, pp. 3-6, 2011.
6. Cahyaningtyas, C., “Speech Synthesis Bahasa Indonesia Berbasis Hidden Markov Model (HMM) pada Intonasi Kalimat Berita dan Kalimat Tanya,” Tugas Akhir Jurusan Teknik Fisika FTI-ITS, Surabaya. Indonesia, 2015.
7. Zen, H., Toda, T., Nakamura, M., dan Tokuda, K., “*Details of Nitech HMM-Based Speech Synthesis*

System for the Blizzard Challenge 2005". IEICE Trans. INF & SYST, Vol. E90-D, No. 1, 2007.
8. Kawahara, H., Katayose, H., Cheveignė, A., and Patterson, R., "*Fixed Point Analysis of Frequency Mapping for Accurate Estimation of F_0 and Periodicity*," Proc. Eurospeech, pp. 2781-2784, 1999
9. Tohkura Y. A., "weighted cepstral distance measure for speech recognition," IEEE Trans Acoust Speech Signal Processing 1986; 35(10): 761–764.