F0-based normalization scheme for MFCC speech features

Włodzimierz Kasprzak

Institute of Control and Computation Engineering

Warsaw University of Technology

INTRODUCTION

Common approach to make speech features speaker-independent:

- to perform a **linear** or **piecewise** linear warping of the frequency axis.
- the warping function is estimated by a ML approach.

Disadvantage of common approach:

- requires a large training material to be collected for a speaker.

On-line approach to speaker-independent feature normalization:

- to explore the locations of main spectral formants.

Observation:
the **basic speech frequency F0** plays a central role in human speech generation.

1. MFCC

The MFCC features are computed according to the following standard equations:

1) The *M/2* Fourier power coefficients, k=1,...,M/2; for every window τ under a Hamming window function w_τ(t):

$$FC(k,\tau) = |\ F(k,\tau)\ |^2 = \left|\ \frac{1}{M}\sum_{t=0}^{M-1}[x(\tau+t)e^{-i2\pi kt/M}\cdot w_\tau(t)]\ \right|^2$$

2) The Mel-spectral coefficients from L triangle filters $D(l)$:

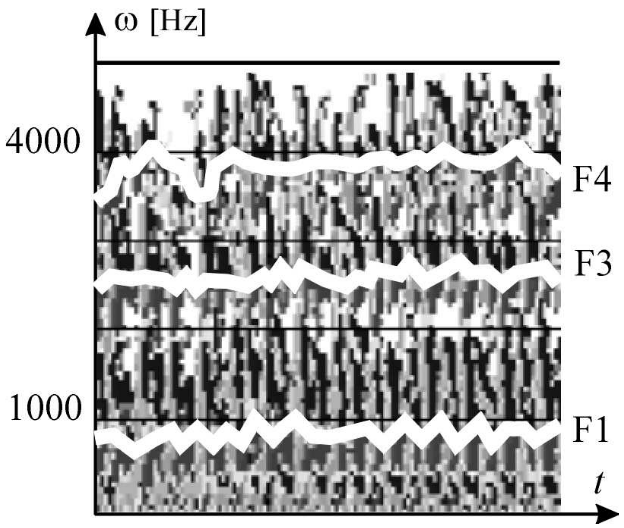
$$MFC(l,\tau) = \sum_{k=0}^{M-1}[D(l,k)\cdot FC(k,\tau)].$$

distributed according to the **Mel scale**.

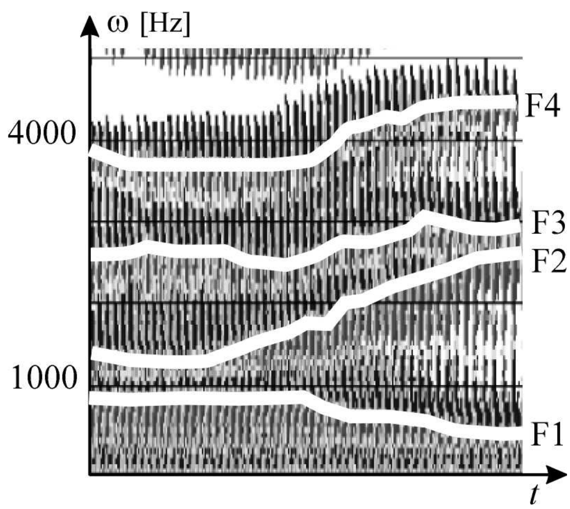
3) The MFCC coefficients, k=1,...,12,

$$MFCC(k,\tau) = \sum_{l=0}^{L-1}[\log MFC(l,\tau)\cdot\cos(\frac{k\cdot(2l+1)\pi}{2L})]$$

2. Phonemes

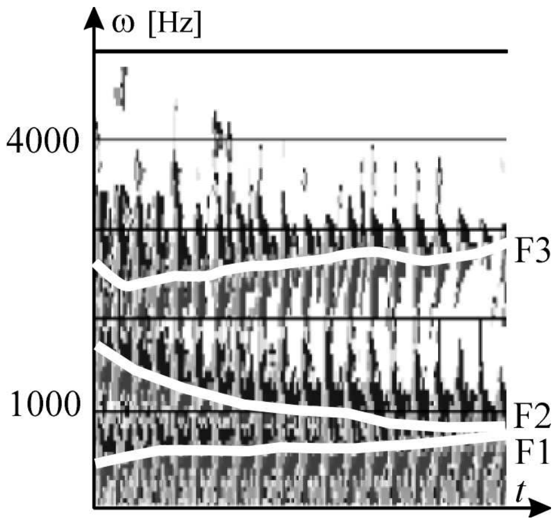


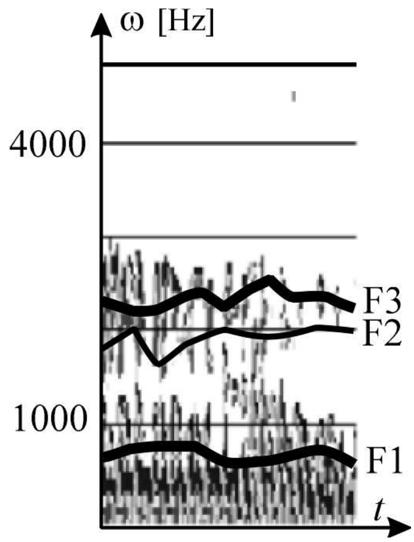
Monophtong /a/.



Diphtong /aI/

2. PHONEMES

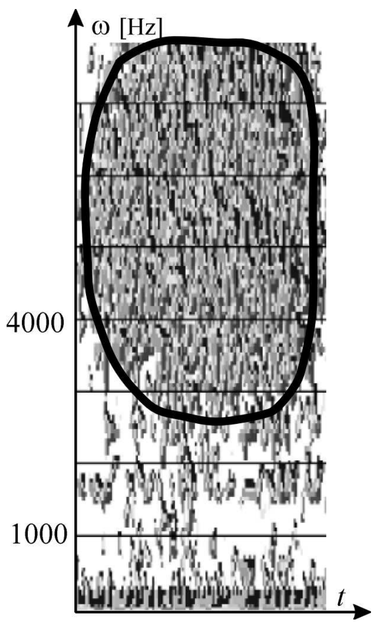




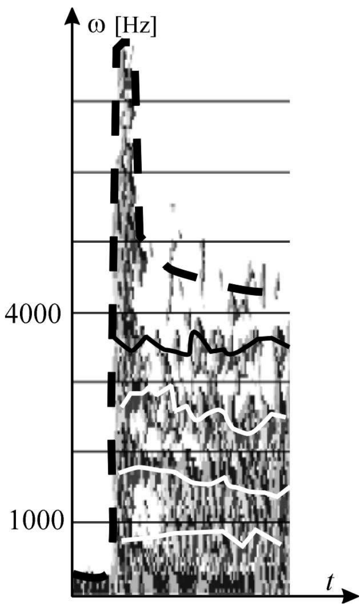
Approximant ("liquid") / *l* /.

Nasal /n/

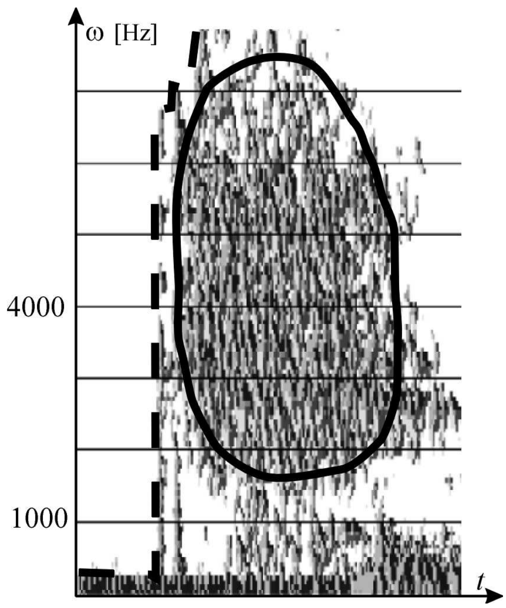
2. PHONEMES



Fricative /s/.



Plosive ./t/



Africate /tS/

.

3. Our approach

A F0-based frequency normalization scheme will be applied **selectively** to some categories of phonemes only, as MFCC already do not locate spectral peaks precisely.

General phoneme **selection criteria**:

- voiced – unvoiced,
- vowels and strong consosnants vs. weak consonants.

Our approach:

1) Detecting the current F0
2) Determining current phoneme type
3) Performing a correction in the Mel frequency scale.

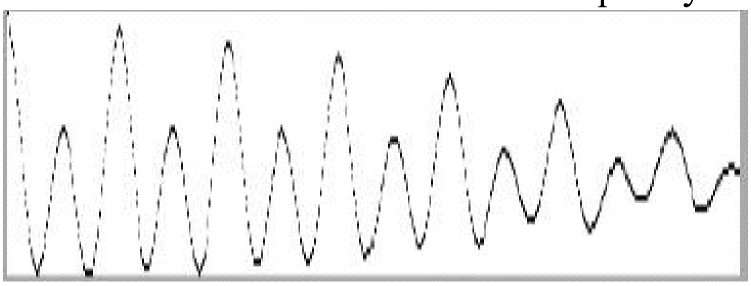
3. Our approach

1) Detecting the current F0

We compute the signal's auto-correlation in time while using an enlarged window of double size, compared to one used for the Fourier transform:

$$r_k^{(\tau)} = \frac{\sum_{n=\tau}^{\tau+N-k-1} f_n f_{n+k}}{|[f_n]| |[f_{n+k}]|_n}, \; k=1,..,N$$

The non-zero index k, for which $r_k^{(\tau)}$ attains a maximum, together with the sampling frequency, determines the instantaneous basic frequency of the signal.

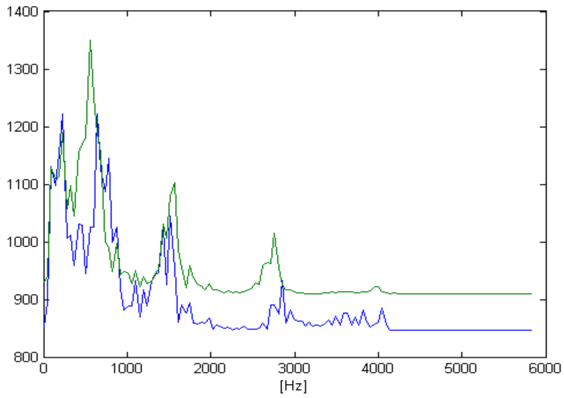


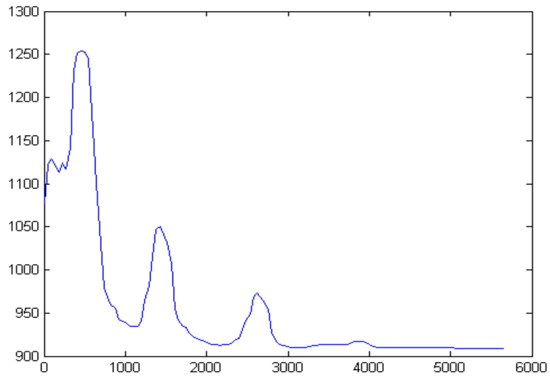
The auto-correlation function $r_k^{(\tau)}$, for a fixed τ, for a vowel.

3. OUR APPROACH

2) Determining current phoneme type

Selecting between vowels and other phonemes is due to the ananlysis of eneregy distribution in given window.





Example of two distributions of FC coefficients (log-scale) for vowel /e/: directly (left drawing) and after 5 point-smoothing (right drawing).

3. OUR APPROACH

3) Performing a correction in the Mel frequency scale.

The mapping between an original Mel frequency value f_{MEL} and the corrected Mel frequency f_{new} depends on the difference of currently measured f_{F0} and the assumed normalized frequency $f_{\mathrm{F0\text{-}norm}}$:

$$f_{MELnew} = f_{MEL} + \kappa \cdot (f_{\mathrm{F0}} - f_{\mathrm{F0\text{-}norm}})$$

For voiced phonemes the correction parameter κ is set to 0.6, and for unvoiced phonemes κ = 0.1.

4. EXPERIMENTS

The goal of experiments was to evaluate the similarity of MFCC feature sets for every individual phoneme category, before and after a F0-based correction.

The **qualty of features** for a single phoneme is expressed by two **error measures**.

1) To represent the **average within-class** or **between-class square distance** for feature vectors: Let S_N be a set of N feature vectors with L features each. S_N is compared with S_M, that contains (in general) M vectors. The average matching error for these two sets is:

$$\varepsilon(S_M, S_N) = \frac{1}{M \cdot N} \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{1}{L} \sum_{l=1}^{L} (x_l^n - x_l^m)^2$$

2) The total within-class variance of all features from both sets is:

$$\sigma^2(S_M, S_N) = \frac{1}{M+N} \sum_{j=1}^{M+N} \frac{1}{L} \sum_{l=1}^{L} (x_l^j - \bar{x}_l)^2$$

4. EXPERIMENTS

Sample set

1) 200 word utterances, coming from 4 speakers (2 male and 2 female speakers),
2) manually labeled
3) 12 selected phonemes:
   - vowels /a/, /e/, /o/;
   - approximants /y/, /r/,
   - nasal /n/,
   - fricatives /z/,/v/;
   - affricates /dZ/, /tS/; and
   - plosives /t/,/d/.

4. Experiments

1) Similarity of features before correction

Comparison of feature vector distances and variances within the same speaker for different phonemes.

Speaker	Average distance ε within-class	Total within-class variance σ^2	Average between-class distance ε	Average F0 [Hz]
Vowel /a/				
Male 1	2.22	0.76	19.2	118
Male 2	2.39	0.76	16.5	125
Female 1	2.42	0.77	19.5	188
Female 2	2.34	0.78	21.2	206
Average	**2.28**	**0.77**	**19.1**	
Vowel /e/				
M1	1.13	0.55	13.5	119

4. Experiments

M2	1.91	0.71	13.4	136
F1	1.78	0.68	14.1	201
F2	1.91	0.71	19.2	199
Average	**1.52**	**0.63**	**15.1**	
Vowel /o/				
M1	2.65	0.83	18.9	119
M2	5.81	1.08	17.6	132
F1	3.04	0.88	20.4	193
F2	2.58	0.82	23.4	183
Average	**2.62**	**0.83**	**20.1**	
Approximant /y/				
Average	**3.35**	**0.92**	**18.5**	
Approximant /r/				
Average	5.56	1.05	19.0	
Nasal /n/				

4. EXPERIMENTS

Average	3.85	0.94	21.8	
Fricative /z/				
Average	7.21	1.12	29.4	
Fricative /v/				
Average	3.93	0.84	14.7	
Affricate /dZ/				
Average	8.24	1.05	23.1	
Affricate /tS/				
Average	3.20	0.66	40.5	
Plosive /t/				
Average	1.93	0.48	14.3	
Plosive /d/				
Average	1.82	0.63	13.8	

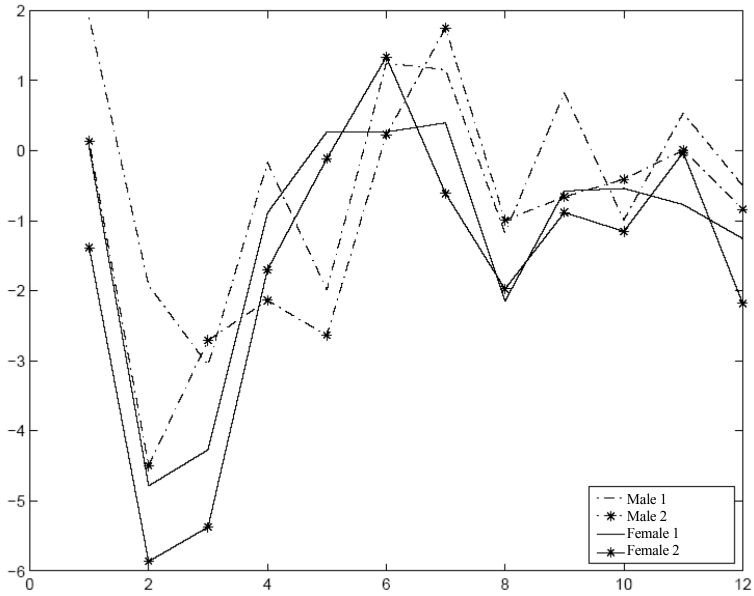
4. EXPERIMENTS

2) Summary of observations

- **Best quality** (small within-phoneme distances and variances, and large between-phoneme distances) shown by **plosives** and **unvoiced affricates**.
- The between-phoneme distances are sufficiently large if compared to within-phoneme distances for every single speaker.
- For a **vowel** and a **nasal** strong differences between speakers exist.
- An **affricate** and **consonant** show better between-speaker similarities than a vowel and nasal.

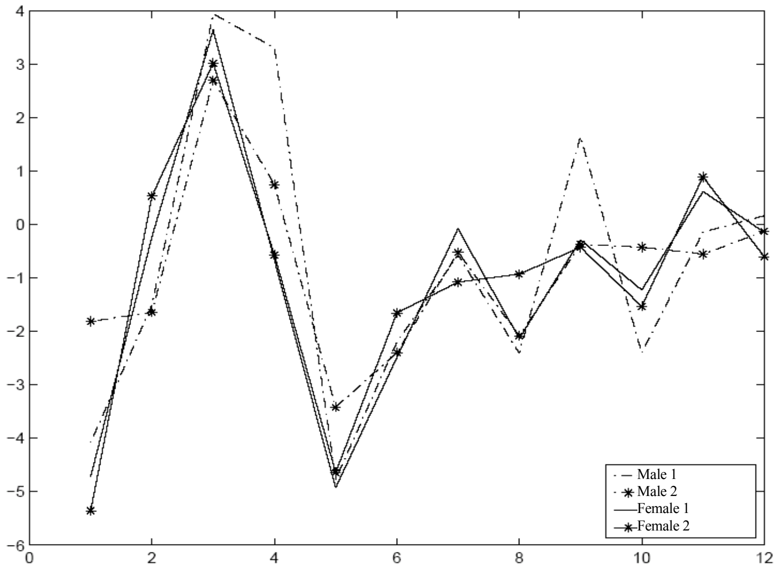
- A correction procedure should operate differently according to phoneme categories.

4. Experiments



Average MFCC feature vectors for vowel /a/ for 4 speakers.

4. Experiments



Average MFCC feature vectors for approximant /y/ for 4 speakers.

4. EXPERIMENTS

3) SIMILARITY OF FEATURES AFTER CORRECTION

Our feature correction experiments process only male or only female utterances separately - the $f_{F0\text{-}norm}$ is set to 200 Hz or 130 Hz, accordingly.

Our scheme improves the stability of MFCC features generated for **vowels** (e.g. /a/, /e/, /o/) and **voiced consonants** (e.g. /y/). The articulation of these phonemes is not disturbed and attenuated by the vocal tract.

In contrast, the **retroflex** /r/ is heavily attenuated and the correction scheme offers no improvement for it.

A positive influence onto the feature stability can also be observed for **affricates**, **especially voiced** ones (e.g. /dZ/).

Unvoiced affricates, like /tS/, have a very stable original features, that seems not to depend from F0.

4. EXPERIMENTS

For **fricatives** the improvement of the correction scheme is visible for **open, non-attenuated phoneme** like /z/, but is not visible and even deteriorates the feature vector for **attenuated phoneme** /v/.

4. EXPERIMENTS

4) PARAMETER SETTING

The experiments with various settings of parameter κ leaded to the conclusion that for voiced phonemes it should be (κ = 0,6) and for unvoiced one - κ = 0,1 - or eventually the correction can be omitted.

TOTAL RESULTS OF CORRECTING FEMALE UTTERANCES BY NORMALIZATION TO $F_{F0\text{-}NORM}$ = 130HZ.

Phoneme	Decrease of distance ε		Decrease of variance σ^2	
	κ = 1.0	κ = 0.6	κ = 1.0	κ = 0.6
Vowels				
/a/	18.1 %	14.8 %	10.7 %	8.5 %
/e/	15.5 %	16.6 %	9.8 %	9.6 %
/o/	4.9 %	9.6 %	2.8 %	4.9 %
Approximants				

4. EXPERIMENTS

/y/	15.7 %	17.6 %	10.0 %	10.1%
/r/	-2.9%	0.9 %	0.2%	2.9%
Affricates				
/tS/	3.7 %	4.5 %	2.9 %	3.8 %
/dZ/	30.4 %	27.5 %	17.6 %	16.9 %
Nasal				
/n/	-6.6 %	-5.6 %	-0.3 %	-1.4 %
Fricatives				
/v/	-7.5 %	-6.9 %	-3.4 %	-1.6 %
/z/	9.1%	5.4 %	5.8 %	4.2 %
Plosives				
/d/	-5.0 %	4.0 %	-5.3 %	0.8 %
/t/	14.0 %	12.6 %	6.3 %	7.1 %
AVERAGE	7.7 %	8.6 %	4.7 %	5.4 %

4. EXPERIMENTS

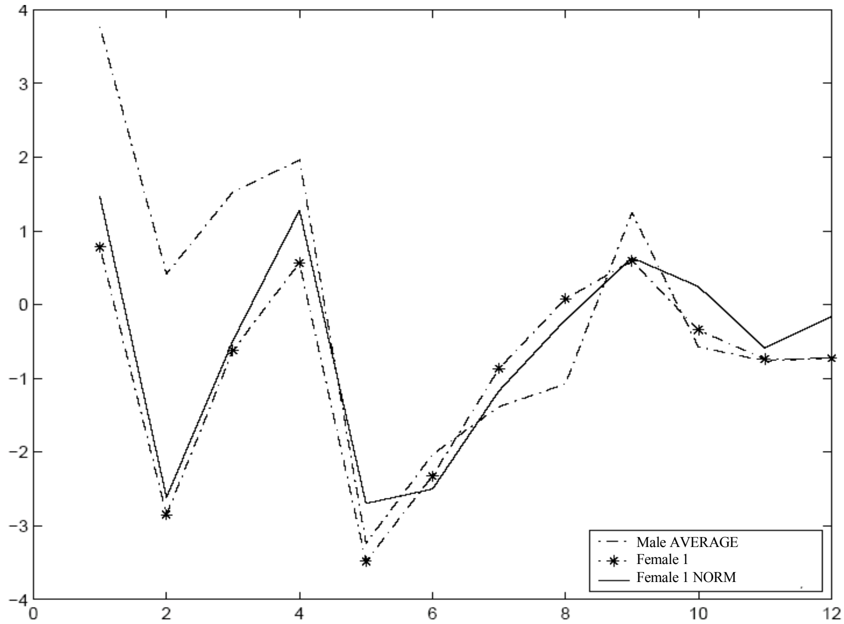
TOTAL RESULTS OF CORRECTING MALE UTTERANCES BY NORMALIZATION TO $F_{F0\text{-}NORM}$ = 200 Hz.

Phoneme	**Decrease of distance ε**		**Decrease of variance σ^2**	
	κ = 1.0	κ = 0.6	κ = 1.0	κ = 1.0
Vowels				
/a/	16.9 %	1.4 %	9.8 %	4.2 %
/e/	15.5 %	16.6 %	9.8 %	9.6 %
/o/	4.9 %	9.6 %	2.8 %	4.9 %
Approximants				
/y/	15.7 %	17.6 %	10.0 %	10.1%
/r/	-2.9%	0.9 %	0.2%	2.9%
Affricates				
/tS/	3.7 %	4.5 %	2.9 %	3.8 %
/dZ/	30.4 %	27.5 %	17.6 %	16.9 %

4. EXPERIMENTS

Nasal				
/n/	-6.6 %	-5.6 %	-0.3 %	-1.4 %
Fricative				
/v/	-7.5 %	-6.9 %	-3.4 %	-1.6 %
/z/	9.1%	5.4 %	5.8 %	4.2 %
Plosive				
/d/	-5.0 %	4.0 %	-5.3 %	0.8 %
/t/	14.0 %	12.6 %	6.3 %	7.1 %
AVERAGE	7.7 %	8.6 %	4.7 %	5.4 %

4. Experiments



Normalizing female speech (with $\kappa = 0.6$) for approximant /y/ to basic frequency of male speech.

5. Conclusion

An on-line correction scheme for MFCC features - that computes the instantaneous F0 frequency of speech and accounts for different phoneme categories.

The original MFCC features are **relatively stable** for unvoiced and heavily attenuated phonemes, hence for such phonemes a correction scheme is not necessary.

For **open** or **voiced** phonemes the proposed correction scheme was experimentally verified to decrease the relative error by more than 10%.