THE STATISTICS AND PHONE ERROR RATES OF BARK-FEATURES

Harald Höge

Universität der Bundeswehr München
harald.hoege@t-online.de

Abstract: We simulate the process of feature processing as assumed to be done in the human brain. The simulation is based on the principle that the features are processed **independently** in critical bands. 30 critical bands are realized by a Gammatone filterbank. The output of each band is segmented into phones. From each segment and each band a 'modulation feature vector' is extracted assuming that the spectrum of modulation is stationary during the duration of a phone. Using GMMs trained on those modulation features, a recognizer is constructed for each filter. The segments are classified[1] into phonemes leading to a phone error rate per band. Given the emission probabilities of the GMMs the probabilities for each phone and for each band are determined. In our approach these probabilities build the components of a 'phone feature vector', which is assumed to be processed in the auditory cortex. To the authors knowledge the transformation of the modulation features to phone features is unknown neuro-physically. Yet from perceptive experiments we know some statistic properties of the phone features concerning the relation between the human phone error rate per band and the human error rate of unfiltered phones [2]. To evaluate this relation we combine the 30 phone feature vectors and construct an 'all-band' recognizer. Classification of segments of Spanish speech using 32 phones leads to an 'all band' phone error rate of 48% and a phone error rate per band of about 92%. These error rates deviate significantly from human performance.

1 Introduction

The features extracted for human phone perception are generated along the auditory pathway by several transformations [1]. The pathway is built up by three subsystems: the cochlea located in the inner ear, the inferior colliculus located in the brainstem, and the auditory cortex. On this pathway the information given by the vibration of the basilar membrane undergoes several complex transformations. The first transformation is performed by the inner hair cells sampling the vibrations of the basilar membrane [3]. The function of each hair cell can be described by a band-limited filter, where the output of each filter is rectified and smoothed. The resulting information of the hair cells can be interpreted as a kind of short term spectrum $y(F,t)$ - the auditory signal. For each ear the auditory signal is transported via the cochlea nerves to its cochlea nucleus. At this stage bundles of nerves are evolving, where each bundle transports the information provided by the vibration of about 1 mm along the basilar membrane [1]. Perceptive experiments reflect the processing in bundles and leads to the definition of the Barks scale [9]. We call this information stemming from those '1 Bark bundles '**Bark feature vector (BF_k)**. This information is related perceptively also to a band of frequencies, which is called a **critical band** [9]. In the following we take also this naming instead of Bark. The human brain processes 35 BFs (k=1,…,35 Bark), where about 30 BFs contain information for phone perception. The components of the BFs are outputs of bundles of specific neurons organized in lamina. The information contained in the Bark feature vector changes during the pathway, but this information stems always from the bundle of hair cells sensing the vibration of the basilar membrane in the range of 1 mm. This kind of processing is common to all mammals. In the cochlea the components of BF_k is given by samples of the auditory signal $y(F_i,t)$, where the sampled frequencies F_i belong to a critical band. The first transformation is done in the olive complex containing spatial information, which is neglected in the paper. The next transformation of the BFs is performed in the central inferior colliculus.

[1] Classification is defined, when the boundaries of the segments are known.

Here neurons are tuned to specific frequencies of amplitude/frequency modulation. We call the information contained in the BFs at this stage '**modulation feature vectors' (MF_k)**. It should be emphasized that the MFs are not tuned to phone perception. The last transformation of the Bark feature vectors occur in the auditory cortex. We call the information contained in the BFs at this last stage '**phone feature vectors' (PF_k)** [2]. The neurons providing the components of the phone feature vectors are tuned in the age of less than 1 year of a child to the phones of its language [12]. In this age the speed of learning is very fast. With increasing age the speed of learning decreases rapidly. To the author's knowledge the transformation of the modulation feature vectors to the phone feature vector is not known. As shown in Figure 1 we speculate that each phone feature vector is used in the secondary auditory cortex to perceive phones bandlimited to the corresponding critical band (band perception). We define phone error rates e_k k=1,…,K measured for each critical band. To perceive phones from all critical bands the phone feature vectors are combined (all bands perception).

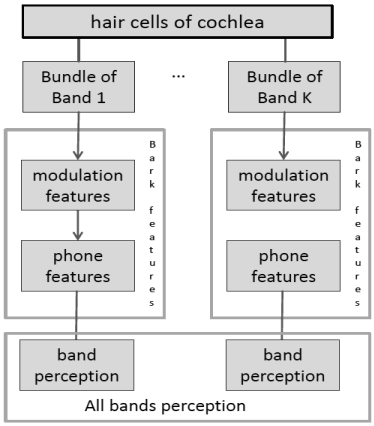

Figure 1 – System Architecture of processing information originating of K bundles of hair cells. Each bundles sample the vibration of the basilar membrane in the range of 1 Bark

Fletcher [4] measured such error rates introducing the concept of articulation filters. Their bandwidth is constructed in such a way, that the phone error rates of phones building nonsense syllables are equal for all filters in unnoisy conditions. It turns out that the bandwidths of the articulatory filters are equal to the bandwidths of the critical bands except for very low and very high frequencies. According to the investigations of Fletcher and Steward [4] there exist an interesting relation between the human phone error rates e_k (PER-band) and the phone error rate e (PER-all-bands) of unfiltered speech (summing up the outputs of all K bands):

$$e = \prod_{k=1}^{K} e_k \qquad (1)$$

This relation must origin from specific statistic property of the phone features, which are unknown.

Our long term goal is to build recognizers for each band and a recognizer for all bands with following properties:

1. The phone error rates should obey the relation (1)
2. The phone error rates e_k should be equal
3. The minimal phone error rate e should be less than 10% [2]

[2] In his experiments Fletcher found a minimal phone error rate of 1.5%. Our investigations are done with conversational speech. For this experimental setting we assume higher minimal human phone error rates. The lowest phone error rate achieved on the TIMIT data base is 15.7% [13].

To investigate, if these properties hold also for the simulated feature, we use as frameworks the architecture shown in figure 1. The focus of the paper is to develop this frame work. The building blocks of the architecture are crude models for the modulation vector and the phone feature vector. The models have to be tuned in future work. The paper is organized as follows:

The auditory signal is generated by a Gammatone filterbank of 30 filters. The filterbank and the construction of the modulation-feature vectors are described in chapter 2. Chapter 3 describes the construction of the phone features and the recognizers. Results on the phone error rates e_k and e are presented in chapter 4.

2 Modulation Features

The modulation features are generated in 2 steps. First the speech signal is filtered by a Gammatone filterbank, which is approximated as described in [7]. The filterbank is described in section 2.1. Section 2.2 describes the construction of the modulation feature vectors MF_k.

2.1 The Gammatone Filterbank

A Gammatone filter is a specific bandpass which is defined by the center frequency F_k and the 3 dB bandwidth f_b. For constructing a Gammatone filterbank, the center frequencies and the bandwidths are chosen based on psychoacoustic measurement [8]. The complex impulse response h_γ of a Gammatone filter of order γ is given by

$h_\gamma(n) = n^{\gamma-1} \cdot a^n; a = \lambda e^{i\beta}$

λ denotes the bandwidth parameter; β denotes the oscillation frequency. The z-transform of the impulse response for a first order Gammatone filter is given by $h_1(z) = \frac{1}{1-az^{-1}}; a = \lambda e^{i\beta}$ which leads to an analytic signal given by the operation $y_n = ay_{n-1} + x_n$. Following [7] the z-transform of a Gammatone filter of any order is approximated by

$g_\gamma(z) = h_1^\gamma(z) = \frac{1}{(1-az^{-1})^\gamma}; a = \lambda e^{i\beta}$ The corresponding filter operations can be done by cascading the first order filter. The value of β is defined by the center frequency F_k and the sampling frequency f_s: $\beta = 2\pi \frac{F_k}{f_s}$

For a 3 dB bandwidth f_b the damping factor λ is given by

$$\lambda = -\frac{p}{2} - \sqrt{\frac{p^2}{4} - 1}; p = \frac{-2 + 2\alpha \cos\left(2\pi \frac{f_b}{2f_s}\right)}{1-\alpha}; \alpha = 10^{-\frac{3}{40}} \quad (2)$$

The center frequencies F_k of the Gammatone filterbank are equally spaced on the ERB scale. The ERB scale is related to the frequency by the relation

$$ERB_{aud}(f) = 24{,}7 + \frac{f}{q}[ERB]; f = 24.7\, q\left(e^{\frac{ERB}{q}} - 1\right); q = 9.265 \quad (3)$$

For a given value ERB the damping factor λ and the 3dB bandwidth f_b is given by [7]

$$\lambda = e^{-\frac{2\pi\, ERB}{a_\gamma f_s}};\ a_\gamma = \frac{\pi(2\gamma-2)!\,2^{-(2\gamma-2)}}{(\gamma-1)!^2}\ f_b = \frac{c_\gamma}{a_\gamma} ERB;\ c_\gamma = 2\sqrt{2^{\frac{1}{\gamma}} - 1} \quad (4)$$

2.2 Modulation Features

The filters F_k, k=1,…,K with center frequencies F_k deliver the analytic signals $y_k(t, \Omega_k)$, k=1,…K. These signals can be represented by [6, pp.369]

$$\left.\begin{aligned} y_k(t, F_k) &= |y_k(t, F_k)| e^{j\psi(t,F_k)} \equiv a(t, F_k) + jb(t, F_k) \\ |y_k(t, F)| &= |a^2(t, F_k) + b^2(t, F_k)|^{\frac{1}{2}};\ \psi(t, F_k) = tan^{-1}\frac{b(t,F_k)}{a(t,F_k)} \end{aligned}\right\} \quad (5)$$

The real part of $y_k(t,F_k)$ can be regarded as an amplitude modulated cosine wave (see example in figure 3) with modulated phase around $2\pi F_k t$:

$y_k(t,F_k) = |y_k(t,F_k)|cos[2\pi F_k t + \theta(t,F_k)]$;

We define $\dot{\psi} \equiv \frac{d}{dt}\psi = 2\pi F_k + \dot{\theta}(t,F_k)$ as the instantaneous frequency and $\dot{\theta}$ as the instantaneous frequency deviation. The phase ψ can be recovered by

$\psi(t,F_k) = \int_0^t \dot{\psi}\,(t,F_k)dt + \psi(0,F_k)$

The instantaneous frequency deviation $\dot{\theta}$ is given by [6, p.370]

$$\dot{\theta}(t,F_k) = \frac{b(t,F_k)\dot{a}(t,F_k) - a(t,F_k)\dot{b}(t,F_k)}{a^2(t,F_k)+b^2(t,F_k)} \quad (6)$$

The signals $y_k(t,F_k)\ k = 1,\dots,K$ are cut into segments with starting time t_1 and end time t_2 representing the signal of a phone. The timeslot $[t_1,t_2]$ is given by the alignment labels of the speech databases. Given $[t_1,t_2]$ of a segment, the AM-spectrum $AM_k(t_1,t_2)$ derived from the amplitude modulation $|y_k(t,F)|$ and the FM-spectrum $FM_k(t_1,t_2)$ derived from the instantaneous frequency deviation $\dot{\theta}(t,F_k)$ is given by their Fourier transform:

$$AM_k(F) = \left|\int_{t_1}^{t_2}|y_k(t,F_k)|e^{jFt}dt\right|;\ FM_k(\Omega) = \left|\int_{t_1}^{t_2}\dot{\theta}(t,F_k)e^{jFt}dt\right| \quad (7)$$

The functions $AM_k(F)$ and $FM_k(F)$ are sampled building the components of the modulation-feature vector BF_k (see chapter 4).

3 Phone Features

For classification we use GMMs trained for each band. For each phone Ph_j the emission probabilities $p(MF_k|Ph_j)$ are approximated by GMMs with tied covariance matrices cov_k

$$p(MF_k|Ph_j) \approx \sum_{\rho=1}^{N_{j,k}} c_\rho N(mean_{jk\rho}, cov_k) \quad (8)$$

We define phone error rates e_k, provided by the classifier based on Bayes decision rule

$$\widehat{Ph}_k = argmax_{Ph_j} p(MF_k|Ph_j)P(Ph_j) \quad (9)$$

This classifier uses a mono-phone language model $P(Ph_j)$. To determine the error rate e of the unfiltered speech signal we use 2 approaches. The first approach assumes that the modulation features MF_k are statistic independent for the different bands. This assumption leads to the emission probability $p(\overrightarrow{MF}|Ph_j) \equiv \prod_{k=1}^{K} p(MF_k|Ph_j)$ and a related classifier:

$$\widehat{Ph} = argmax_{Ph_j} p(\overrightarrow{MF}|Ph_j)\ P(Ph_j), \quad (10)$$

This classifier delivers an all-band phone error rates denoted as e^I. The second approach is in the spirit of the architecture shown in figure 1. Phone-features are derived from $p(MF_k|Ph_j)$. The components of the phone feature vector PF_k are defined by

$$PF_k(j) = -\log P(Ph_j|MF_k);\ j = 1,\dots,nPh$$
$$= -log\frac{p(MF_k|Ph_j)}{p(MF_k)}P(Ph_j);\ p(MF_k) = \sum_{j=1}^{nPh} p(MF_k|Ph_j)P(Ph_j) \quad (11)$$

Following (8) and (9) but using the features (11) a all band classifier is constructed. The resulting all-band phone error rates are denoted as e. In addition we define an all-band phone error rate $e^{I\!I}$, which is derived from (1), where the error rates are e_k are given from the band-classifier (9). The difference between $e^{I\!I}$ and e show the deviation to human perception.

4 Experiments

Our experiments are performed with Spanish speech databases covering broadcast news, conversations and podcast downloaded from various internet sources. The databases were

developed during the QUAERO project and used during for ASR evaluation [11]. The labeling into phone segments was performed by the HMM training system [10] using 3 or 6 state right to left HMMs based on tri-phones. The phones are defined by the acoustic realization of 32 phonemes including the 'non-speech-segment' /si/. The speech signal x(t) is sampled with16kHz providing the samples $x(n)$.

4.1 Gammatone Filterbank

The Gammatone filterbank is constructed as described in 2.1. We use a Gammatone filterbank of 30 filters [7] as specified according to table 1.

filter spec in [Hz]	
sampling rate	16 000
lower frequency	70
base frequency	1000
upper frequency	6700

center frequencies [Hz}				
73.2	348.4	874.3	1879.2	3799.4
107.7	414.2	1000.0	2119.4	4258.5
146.0	487.5	1140.1	2387.1	4770.0
188.7	569.1	1296.1	2685.2	5339.7
236.3	660.1	1469.9	3017.3	5974.4
289.4	761.4	1663.5	3387.3	6681.4

Table 1 – Specification of the Gammatone filterbank with center frequencies F_k k=1,..,30

The real part of the impulse response and its spectrum of a filter is depicted in figure 2.

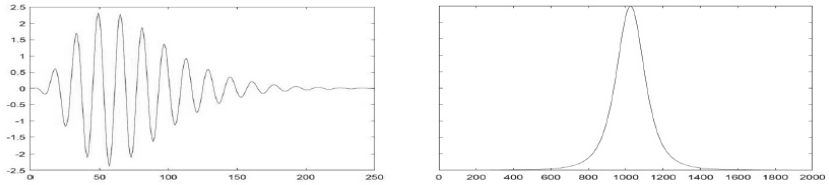

Figure 2 – left: real part of impulse response of filter F_{14} with center frequency of 1000Hz
right: spectrum of impulse response

4.2 Generation of Modulation Features

To calculate the modulation features, equation (5) and (6) have to be evaluated. We regard samples $y_k(n, F_k)$ $n = 1, \dots, N$ of the filter outputs aligned to a time slot $[t_1, t_2]$ of a phone (see section 2.2). The samples $|y_k(n, F_k)|$ deliver the envelopes of $y_k(n, F_k)$ (see figure 3). For implementing (6) the instantaneous frequency deviation $\dot{\theta}$ has to be provided. $\dot{\theta}$ is approximated by first order differences $\Delta a(n) = a(n+1) - a(n); \Delta b(n) = b(n+1) - b(n)$ leading to

$$\Delta\theta(n, F_k) = \frac{b(n, F_k)\Delta a(n, F_k) - a(n, F_k)\Delta b(n, F_k)}{a^2(n, F_k) + b^2(n, F_k)}$$

The FFT of $|y_k(n, \Omega_k)|$ and $\Delta\theta(n, F_k)$ delivers samples of the features of $AM_k(m)$ and $FM_k(m)$ defined in (7). The samples $AM_k(m)$ are normalized by the energy of the segment by following rule:

$$meanBandAM = \frac{1}{N}\sum_{m=1}^{N}|y_k(n, F_k)|\,;\; meanAllBandAM = \frac{1}{N}\sum_{m=1}^{N}|x_k(n, F_k)|$$

$$AM_k(0) \leftarrow \frac{meanBandAM}{meanAllBandAM}; AM_k(m) \leftarrow \frac{AM_k(m)}{meanBandAM}\;; m = 1,2, \dots$$

We use an FFT of order 13 (8192 points). If the number of samples of a phone is smaller than 8192 (corresponds to 510ms) the samples are zero-padded (usual case). Otherwise the signal is cut to 8192 samples (happens often in silence segments). For the sound plotted in figure 3 the modulation spectra are shown in figure 4. The energy of the spectra is concentrated at low

frequencies. In this way each phone is represented by a modulation feature vectors MF_k; k=1,…,30 with dimension $dim_{AM} + dim_{FM}$ given in tables 1 and 2:

$$MF_k = [AM_k, FM_k];\ AM_k(m), m = 1, \dots, dim_{AM}; FM_k(m), m = 1, \dots, dim_{FM}.$$

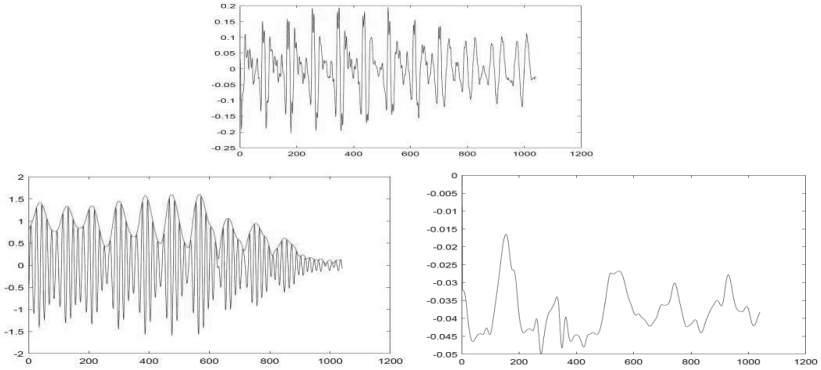

Figure 3 – above: original signal of the sound labeled as /a/ with a duration of 1000 points (62,5ms); left below: real part and the absolute value of the phone /a/ filtered by filter $F_{14}(\Omega_{14} = 1kHz)$; right below: instantaneous frequency deviation

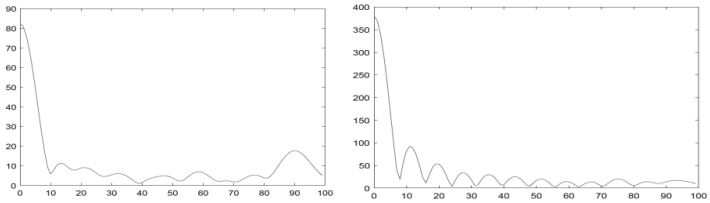

Figure 4 – modulation spectra; left: AM-spectrum in Hz; right: FM-spectrum in Hz.

4.3 Classification

From the speech databases sets of MF_k for each phoneme and each band are generated. Given those sets, a LDA (linear discriminant analysis) is generated, which transforms the feature vectors to lower dimension. Further these sets are used to train 'MF-GMMs'. According to (8) the covariance matrices of each band are tied. The classifier (9) delivers the error rates e_k. The all band error rate of the combined filters is determined for the two cases described by (10) and (11). Equation (11) leads to high dimensional phone feature vectors of dimension $30 \cdot (dim_{AM} + dim_{FM})$. Their dimensionality is decreased by applying an LDA. From these vectors we train 'PF-GMMs', which are used for classification.

4.4 Results

The size of speech databases used for training and testing is given by their number of phones:

test	training
288 028	1 499 809

To answer the question, if the sizes of the databases are sufficient to achieve relevant results, the GMMs and LDAs were trained on the test databases as well as on the training databases. Performing classification on the test database we compared the two phone error rates from the different GMMs and LDAs. We found, that the difference in phone error rates for the 2 cases is about 1% absolutely. Due to the high error rates observed, the magnitude of these deviations can be neglected. Thus the amount of data available are sufficient[3].

[3] Overfitting could happen by the use of the PF-LDA with the high dimension of 960 x 960

Preliminary experiments were done to investigate the discriminative power of the AM_k and FM_k features. As shown in table 2 the discriminative power of the AM-features is higher than those of the FM-features. For this experiment training and test was done on the test databases. For the definition of the phone error rate e^I see chapter 3.

dimAM	dimFM	dimMF	e^I
50	0	20	65,4
0	50	20	82,0

Table 2 – phone error rates in %; dimAM (dimFM) = number of points of amplitude (frequency) modulation spectrum (see figure 4); dimMF=dimension of modulation feature vector after LDA

In the following experiments the AM_k and FM_k features are concatenated building the modulation feature vector MF_k. Training of the MF-GMMs and MF-LDAs are performed on the training databases. In table 3 the dimension of the feature vectors used is shown.

dimAM	dimFM	dimMF	dimPF
50	50	30	32

Table 3 – dimensions of vectors used; dimPF=dimension of the phone feature vector after LDA

The classification results on the test databases are shown in table 4. The biggest change of improvement is achieved, when increasing the number of modes for modeling the distribution of the modulation features. Increasing the number of modes for the probability feature vector leads to far less improvements.

MF-modes	PF-modes	e^I	e^{II}	e
32	32	64,2	18,8	66,7
120	32	65,5	21,7	48,3
120	128	65,5	21,7	47,2

Table 4 – all band error rates; MF (PF)-modes= number of modes used for the modulation (phone) - GMMs;

Table 5 shows the error rates e_k for the experiment, where the lowest value of e is achieved.

Band	1-5	6-10	11-15	16-20	21-25	26-30
	95	92	91	92	93	92
	94	92	91	92	93	92
	93	92	91	92	93	92
	92	91	91	92	93	92
	92	91	91	92	92	92

Table 5 – e_k in % for each band excluding the /si/-phone

5 Conclusion

We have provided an architecture, which mimics the feature extraction along the auditory pathway. We use the concept of modulation features as processed in the central inferior colliculus[4] and the concept of phone features as processed in the auditory cortex. Due to neuro-physical investigations the functionality of the modulation features is well explored. To the author's knowledge this is not the case for the phone features. We assume that the phone features are related to the probability of the phones within each band. These probabilities are given by the emission probabilities of GMMs trained with the modulation features. In order to gain inside in the deviation to human perception, we use (1) and the fact that the human errors e_k are equal for all k for unnoisy speech. Due to Fletcher investigations the human 'all band error rate' is 1.5% measured for nonsense syllables in absence of noise. Using 30 bands, humans would achieve for e_k a value of about 87%. This value has to be compared with the

[4] A more elaborate cordical model can be found in [5]. The concept of the modulation vector is used also in [14].

values for of e_k in table 5, which are about 92%. Due to (1) this value has a big influence on the all-band error rate e. Thus the difference of 4% is this is quite a big gap. The error rates in table 5 are quite equal in the different bands as found by human perception. Comparing the values of e^I and e in table 4, we see that the assumption of statistic independence of the features MF_k is not correct as the phone features deliver lower error rates. Further the statistic properties of the MF_k lead not to (1) as the error rates e and e^{Π} are not equal. From these results we make 2 conclusions

- We have to find better modulation features to close the gap between the human e_k of 87% and the achieved e_k of 92%
- We have to find better phone features, which follow the product rule (in our case to close the gap between 22% and 47%.

The issue behind the first conclusion can be attacked by modeling the non-stationary character of modulation. The issue behind second conclusion is much harder, as to the author's knowledge a suited theory combining the statistics of features properties to error rates is missing.

References

[1] Winer, J. A., Schreiner, C. E.: The Inferior Colliculus. Springer Verlag, 2005

[2] Höge, H.: On the Nature of the Features Generated in the Human Auditory Pathway for Phone Recognition. Interspeech Dresden: 2015

[3] Dau, T. Püschel, D., Kohlrauch, A.: A quantitative model of the "effective" signal processing in the auditory system. I. Model structure, J. Acoust. Soc. Am. 99 (6), June 1996:pp. 3615:3622

[4] Fletcher, H., Galt, R.H.: The perception of Speech and Its Relation to Telephony. The Journal of the Acoustic Society of America, Vol. 22, number 2: 1950, pp. 89-151

[5] Chi, T., Ru, P., Shamma, S. A.: Multiresolution spectrotemporal analysis of complex sounds: J. Acoust. Soc. Am. 118, August 2005, pp. 887–906.

[6] Quatieri, T. F.: Discrete-Time Speech Signal Processing. Prentice Hall Signal Processing Series, Upper Saddle River, NJ07458. 2001

[7] Hohmann,V.: Frequency analysis and synthesis using a Gammatone filterbank. ACTA ACUSTICA UNITED WITH ACUSTICA, Vol 88, 2002, pp.433-442.

[8] Glasberg, B. R., Moore, B. C. J.: derivation of auditory filter shapes from notched-noise data. Hear. Res. Vol 47, 1990 pp.103-138

[9] Zwicker, E., Fastl, H.: Psychoacoustics, *Berlin: Springer* 1999

[10] Rybach, D., Gollan, C., Heigold, G., Hoffmeister, B., Lööf, J., Schlüter, R., and Ney, H.: The RWTH Aachen University open source speech recognition system. in Proc. Interspeech, Brighton, U.K.: 2000, pp. 2111–2114.

[11] Sundermeyer, M., Nußbaum-Thom, M., Wiesler, S., Plahl, C., El-Desoky Mousa, C.A., Hahn, S., Nolden, D., Schlüter, R., Ney, H.: The RWTH 2010 QUAERO ASR evaluation system for English, French, and German. In Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Prague, Czech Republic:2011, pp.2212–2215.

[12] Kuhl, P., K.: Brain Mechanisms in Early Language Acquisition. Neuron, 67(5): September 2010, pp. 713–727

[13] L.Toth, L.: Combining time- and frequency-domain convolution in convolutional neural network-based phone recognition. ICASSP: 2014, pp. 190-194

[14] S.K.Nemala, Patil, K., Elhilali, M.: A Multistream Feature Framework Based on Bandpass Modulation Filtering for Robust Speech Recognition. IEEE Transactions on Audio and Language Processing, Vol.21: 2013, pp. 416-426,