Rekonstrukcija govora iz MFCC-a u prisustvu Babble šuma

Zoran Veličković, Zoran Milivojević, Danijela Aleksić

***Sadržaj* — U ovom radu određene su performanse MFCC algoritma u prisustvu Babble šuma. Performanse algoritma u uslovima promenljivog SNR-a su određene na bazi MSE rekonstruisanog govornog signala za različite dužine MFCC vektora. Rekonstrukcija govornog signala je obavljena sinusnim modelom zasnovanom na fundamentalnoj frekvenciji govornog frejma i njegovoj aproksimiranoj amplitudskoj karakteristici. Analiziran je uticaj aditivnih Babble smetnji na MSE nastalih od jednog, dva, osam i šezdeset četiri govornika. Komparativnom analizom dobijenih rezultata za MSE određene su performanse MFCC algoritma u uslovima Babble šuma.**

***Ključne reči* — Babble, MFCC, MSE, sinusni model.**

I. UVOD

SAVREMENE aplikacije mobilnih komunikacija zahtevaju pristup ASR (engl. *Automatized Speech Recognition*) sistemima. ASR sistemi su našli primenu u IVR (engl. *interactive voice response*) aplikacijama, servisima u bankarskom poslovanju i u automatizovanim sistemima rezervacije. Tradicionalan način za realizaciju ovih servisa je da se na strani prijema, iz vremenskog oblika govornog signala, izvrši prepoznavanje govora. U [1] i [2] se favorizuje prepoznavanje govora direktno iz kodiranog govornog signala. Glavni nedostatak ovog pristupa je što svaki kodek koristi različite algoritme. Nasuprot sistemima baziranim na kodeku razvijeni su DSR (engl. *Distributed Speech Recognition*) sistemi kod kojih se parametri prepoznavanja izračunavaju na predajnoj strani i šalju prijemnoj strani u RFV-u (engl. *Recognition Feature Vector*) [3]. Na prijemnoj strani RFV se direktno koristi za prepoznavanje govora. Savremeni mobilni servisi pored prepoznavanja govora, često zahtevaju i rekonstrukciju govornog signala iz RFV-a. Najčešće korišćen algoritam za ekstrakciju RFV-a je zasnovan na estimaciji MFCC (engl. *Mel Frequency Cepstral Coefficients*) koeficijenata za frejm govornog signala dužine M. Izračunavanje MFCC koeficijenata se bazira na implementaciji banke filtera sa trougaonom amplitudskom karakteristikom [4]. Banka filtera je komponovana od *p* linearno i *q* logaritamski razmaknutih filtera duž frekventne ose. U literaturi su publikovane varijacije MFCC algoritma u zavisnosti od broja, širine spektara i centralne frekvencije trougaonih filtara [5]. MFCC koeficijenti su razvijeni za potrebe prepoznavanja govora i opisuju specifične karakteristike vokalnog trakta govornika u frekvencijskom domenu. Rekonstrukcija govornog signala iz MFCC koeficijenata je otežana jer u procesu izračunavanja MFCC koeficijenata dolazi do gubljenja detaljnih informacija o spektru. Kod DSR sistema rekonstrukcija govornog signala se realizuje iz aproksimirane amplitudske karakteristike za frejm govornog signala. Aproksimirana amplitudska karakteristika se dobija iz MFCC vektora i na osnovu nje se formira sinusni model govornog signala [6]. Za realizaciju sinusnog modela neophodno je poznavanje vrednosti fundamentalne frekvencije govornog frejma F_0, [7]. Fundamentalna frekvencija se izračunava na predajnoj strani i sastavni je deo RFV-a. Pored podataka o numeričkoj vrednosti fundamentalne frekvencije u sastavu RFV-a šalje se i podatak o tome da li pripadajući frejm sadrži ili ne sadrži govor – (engl. *voicing*). Algoritmi za karakterizaciju frejma opisani su u radu [8]. ETSI Aurora standardom [9] je definisano da se fundamantalna frekvencija F_0 i informacija o karakteru frejma pridružuje RFV-u na predajnoj strani. U [10] je pokazano da performanse MFCC algoritma ne degradiraju značajno u prisustvu WGN-a (engl. *White Gaussian Noise*), dok je u [11] pokazano posredstvom MSE-a (engl. *Mean Square Error*) i MOS (engl. *Mean Opinion Score*) testa da optimalni broj MFCC koeficijenata za rekonstrukciju govornog signala iznosi I=23 za velike vrednosti SNR-a (SNR>20).

U ovom radu izvršena je analiza i određene su performanse MFCC algoritma u prisustvu aditivnog Babble šuma. Babble šum se karakteriše brojem govornika N od kojih potiče. Razmatrana je analiza uticaja Babble šuma koji potiče od jednog (N=1), od dva (N=2), od osam (N=8) i od šezdeset četiri govornika (N=64). Performanse MFCC algoritma u prisustvu Babble šuma određene su na osnovu MSE-a rekonstruisanog govornog signala. Performanse algoritma su određene za karakteristične dužine MFCC vektora I=15 ÷ 25.

Organizacija rada je sledeća. U sekciji 2 je prikazan sinusni model govornog signala i algoritam rekonstrukcije govora iz RFV-a. Vrednosti MSE rekonstruisanog govornog signala u prisustvu Babble šuma prikazane su u sekciji 3. U sekciji 4 izvršena je komparativna analiza dobijenih rezultata.

Z. S. Veličković, Visoka tehnička škola Niš, Aleksandra Medvedeva 14, Srbija (phone: +381-18-588-211; fax: +381-18-588-210; e-mail: zorvel@bankerinter.net).

Z. M. Milivojević, Visoka tehnička škola Niš, Aleksandra Medvedeva 14, Srbija (e-mail: milivojevic@bankerinter.net).

D. A. Aleksić, Visoka tehnička škola Niš, Aleksandra Medvedeva 14, Srbija

II. SINUSNI MODEL

Govor se može predstaviti konvolucijom između pobudnog signala i impulsnog odziva filtra koji predstavlja vokalni trakt govornika [12]. Da bi se opisala čovečija percepcija zvuka govorni signal se filtrira mel-filter bankom a dobijene spektralne komponente se logaritmuju. Spektralna rezolucija dobijena na ovaj način direktno zavisi od broja filtara u mel-filtarskoj banci. Primenom DCT-a na dobijene spektralne komponente određuju se MFCC koeficijenti. Prilikom izračunavanja MFCC koeficijenata gubi se fina struktura spektra uz zadržavanje anvelope amplitudske karakteristike govornog frejma. Ove osobine algoritma za izračunavanje MFCC koeficijenata su korisne u sistemima za prepoznavanje govora ali su u sistemima za rekonstrukciju govora nepoželjne. Algoritam izračunavanja MFCC koeficijenata govornog signala je detaljno opisan u [3], [4].

Postoje dve tehnike za rekonstrukciju govornog signala: SF (engl. *Source-Filter*) i SM (engl. *Sinusoidal Model*). SF model se bazira na filtriranju pobudnih impulsa filtrom koji predstavlja model vokalnog trakta govornika. Frekvencija pobudnih impulsa određena je fundamentalnom frekvencijom F_0 govornog frejma. Rekonstrukcija govornog signala sinusnim modelom realizuje se prema sledećem izrazu:

$$\hat{s}(n)=\sum_{j=0}^{J-1} A_j \cos\left(2\pi\, f_j\, n+\theta_j\right),\ n=0,1,...,M-1, \quad (1)$$

gde je A_j amplituda, f_j frekvencija i θ_j faza j-te spektralne komponente rekonstruisanog signala $\hat{s}$. J prestavlja red sinusnog modela a M dužinu signala. Sinusni model se primenjuje samo kod frejmova koji su selektovani kao govorni. Kvalitet rekonstruisanog signala, pored ostalog, zavisi od tačnosti određivanja fundamentalne frekvencije F_0. Predložen je veliki broj algoritama za određivanje fundamentalne frekvencije F_0 u vremenskom i frekvencijskom domenu [13]. Preciznost određivanja fundamentalne frekvencije F_0 direktno zavisi od frekvencije semplovanja F_S i dužine FFT sekvence. Povećanje preciznosti izračunavanja bazira se na primeni PCC interpolacionog algoritma [14]. Kvalitet rekonstruisanog signala sinusnim modelom, pored preciznosti određivanja fundamentalne frekvencije, zavisi od preciznosti rekonstruisane amplitudske karakteristike $\hat{S}$ iz MFCC vektora. Obzirom da se rekonstrukcija amplitudske karakteristike obavlja na osnovu relativno malog broja elemenata iz MFCC vektora, dolazi do formiranja aproksimirane amplitudske karakteristike koja se razlikuje od originalne. Razlika se prvenstveno ogleda u višim delovima spektra jer se fina spektralna struktura nepovratno gubi u procesu izračunavanja MFCC koeficijenata. Ova razlika dovodi do degradacije rekonstruisanog govornog signala. Izračunavanjem MSE između rekonstruisanog i originalnog govornog signala određene su performanse MFCC algoritma u prisustvu Babble smetnji.

III. EKSPERIMENTALNI REZULTATI

U [10] nisu određene performanse MFCC algoritma za rekonstrukciju govora u odnosu na dužinu MFCC vektora I. U radu [11] je određena optimalna vrednost parametra I u odnosu na kvalitet rekonstruisanog signala za slučaj velikog SNR-a. U ovom radu su određene performanse MFCC algoritma u prisustvu Babble šuma [15] na osnovu MSE rekonstruisanog govornog signala. Performanse algoritma su određene za različite vrednosti parametra I i SNR.

A. Baza govornih signala

Određivanje performansi MFCC algoritma je obavljeno na sledeći način. Prvo je formirana baza govornih signala za *G=5+5=10* govornika (pet muških i pet ženskih) koji su izgovarali iskaz 'Visoka tehnička škola'. Analogni govorni signal iz mikrofona je semplovan frekvencijom F_s*=8kHz* i arhiviran na hard disku čime je formirana baza govornih signala. Baza je proširena signalima koji su generisani dodavanjem Babble šuma (N=1,2,4,8,64). Dalja obrada diskretizovanog govornog signala vrši se MFCC algoritmom koji je implementiran u Matlab-u 7.0. Odabrani su sledeći parametri MFCC algoritma: trajanje frejma *T=25ms* dok preklapanje frejmova iznosi *10ms* čime je određena brzina obrade od 100 frejmova u sekundi. Primenjen je Hamming-ov prozor dužine L=200 sa preklapanjem frejmova od 80 semplova. FFT je dužine N_{FFT}*=512* a broj filtera sa trougaonom amplitudskom karakteristikom u okviru mel-filtarske banke je *K=40* (*p=13*, *q=27*). Dužina DCT-a varirana je u opsegu $I=15\div 25$ što ujedno prestavlja dužinu MFCC vektora.

Rekonstrukcija govornog signala izvršena je na bazi aproksimirane amplitudske karakteristike dobijene za različite vrednosti parametra $I=15\div 25$ i SNR=$-10\div 20$dB. Za rekonstrukciju govornog signala primenjen je sinusni model [13] reda J=15.

B. Srednja kvadratna greška MSE

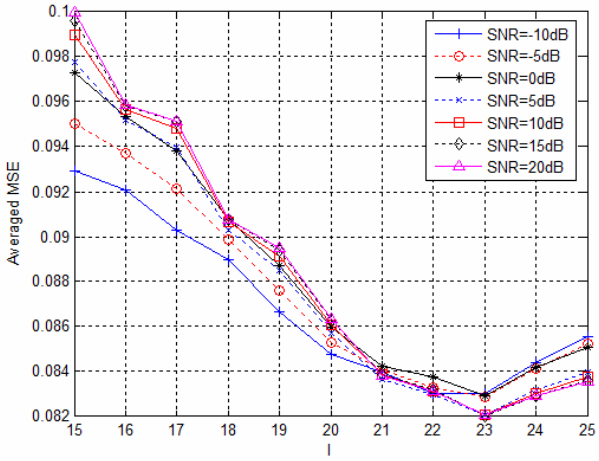
Srednja kvadratna greška MSE određena je na sledeći način:

$$MSE=\frac{1}{M}\sum_{i=0}^{M-1}\left(s(i)-\hat{s}(i)\right)^2 \quad (2)$$

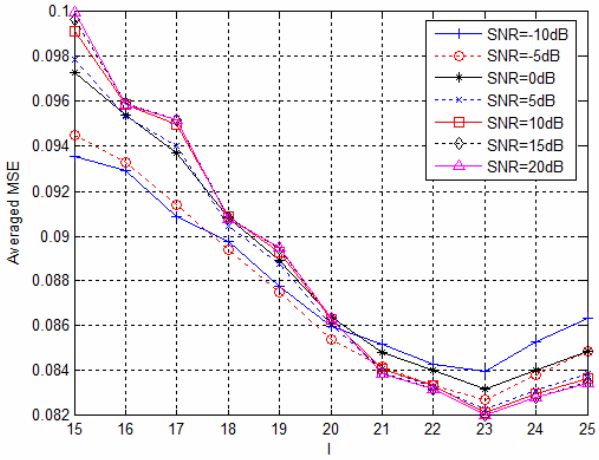
gde je s originalni govorni signal, $\hat{s}$ rekonstruisani govorni signal i M dužina kompletne sekvence govornog signala. Srednja vrednost MSE-a, $\overline{MSE}$, u zavisnosti od dužine MFCC vektora (I) definisana je na sledeći način:

$$\overline{MSE}(I)=\frac{\sum_{g=1}^{G} MSE(I)}{G},\quad I_{\min}\le I\le I_{\max}, \quad (3)$$

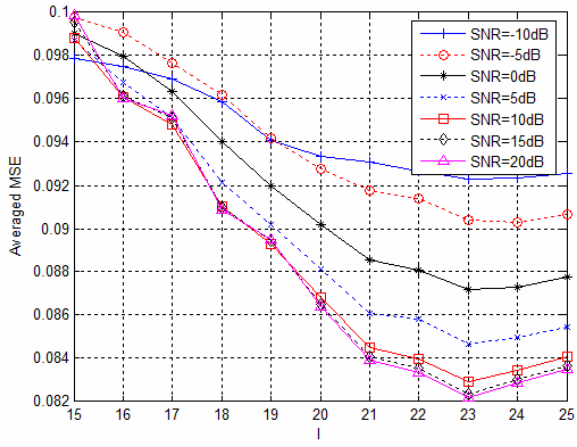
gde je G broj govornika. Srednja vrednost MSE-a, $\overline{MSE}$ (Averaged MSE) je prikazana na slikama sl. 1-4 u funkciji dužine MFCC vektora (I) za karakteristične vrednosti parametra N, sl. 1 (N=1), sl. 2 (N=2), sl. 3 (N=8) i sl. 4 (N=64).



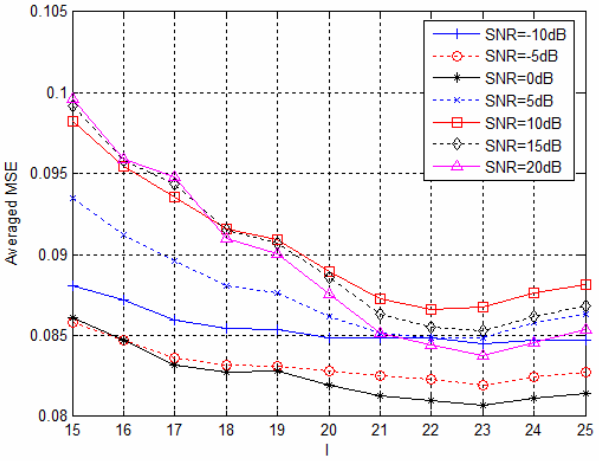
Sl. 1. $\overline{MSE}$ rekonstruisanog govornog signala u funkciji dužine MFCC vektora I za određene vrednosti SNR-a. Babble šum je karakteriziran vrednošću parametra N=1.



Sl. 2. $\overline{MSE}$ rekonstruisanog govornog signala u funkciji dužine MFCC vektora I za određene vrednosti SNR-a. Babble šum je karakteriziran vrednošću parametra N=2.



Sl. 3. $\overline{MSE}$ rekonstruisanog govornog signala u funkciji dužine MFCC vektora I za određene vrednosti SNR-a. Babble šum je karakteriziran vrednošću parametra N=8.



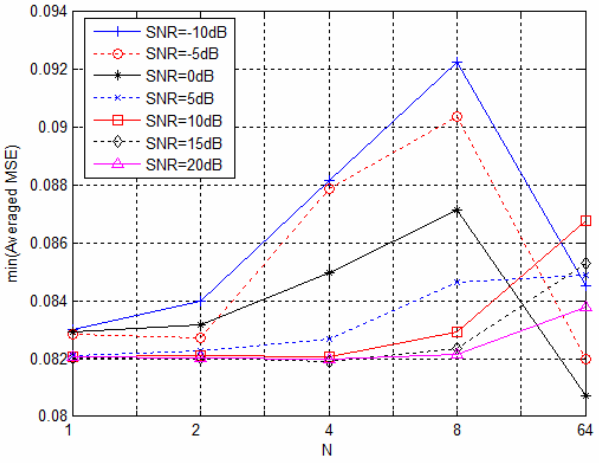
Sl. 4. $\overline{MSE}$ rekonstruisanog govornog signala u funkciji dužine MFCC vektora I za određene vrednosti SNR-a. Babble šum je karakteriziran vrednošću parametra N=64.

U tabeli 1 prikazane su vrednost minimuma $\overline{MSE}$ za različite vrednosti SNR-a i parametara N.

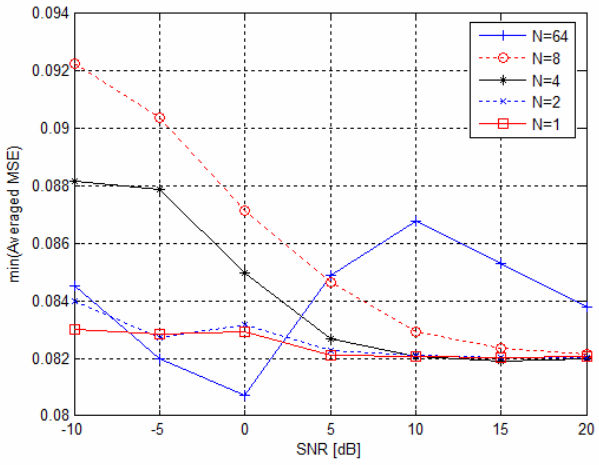
TEBELA 1. VREDNOST MINIMUMA $\overline{MSE}$ U ZAVISNOSTI OD SNR-A I PARAMETRA N.

SNR [dB]	N				
	64	8	4	2	1
-10	0,0845	0,09227	0.08816	0,08397	0,08301
-5	0,08196	0,09039	0.08790	0,08271	0,08286
0	0,08071	0,0872	0.08497	0,08316	0,08294
5	0,08489	0,08463	0.08267	0,08229	0,08210
10	0,08677	0,08294	0.08205	0,08213	0,08205
15	0,08529	0,08236	0.08189	0,08202	0,08204
20	0,08379	0,08216	0.08197	0,08202	0,08206

Na slici 5 prikazana je tranzicija $\overline{MSE}$ u funkciji broja govornika (N), dok je na slici 6 prikazana tranzicija minimuma $\overline{MSE}$ u funkciji SNR-a. Svi minimumi $\overline{MSE}$ su dobijeni za vrednost parametra I=23.



Sl. 5. Tranzicija minimuma $\overline{MSE}$ u funkciji parametra N.



Sl. 6. Tranzicija minimuma $\overline{MSE}$ u funkciji SNR-a.

IV. KOMPARATIVNA ANALIZA REZULTATA

Na osnovu grafika sa sl. 1-6 zaključuje se da:

a. U opsegu parametra I=15÷23 sa povećanjem dužine MFCC vektora dolazi do smanjenja $\overline{MSE}$ za sve vrednosti SNR-a.

b. Minimumi $\overline{MSE}$ dobijaju se za I=23 za sve vrednosti SNR-a. Ovaj zaključak je u saglasnosti sa rezultatima dobijenim u [11].

c. Za N=8 dolazi do velike disperzije minimalne vrednosti $\overline{MSE}$ u funkciji SNR-a i parametra I. Za I=23 i N<8 minimalne vrednosti $\overline{MSE}$ teže vrednosti 0.0823, dok za N=8 postoji veliki raspon $\overline{MSE}$ od 0.082 (SNR=20dB) do 0.0925 (SNR=-10dB). Ovaj zaključak je u saglasnosti sa rezultatima datim u [15] gde se ukazuje na specifičnost Babble smetne za N=8.

d. Kada je parametar $I > 23$ dolazi do blagog povećanja $\overline{MSE}$ nakon čega ($40 > I > 30$) zadržava konstantnu vrednost. Ovaj rezultat je u saglasnosti sa rezultatima dobijenim u [11] na osnovu kojih je određena optimalna vrednost dužine MFCC vektora I=23.

e. Minimum $\overline{MSE}$ raste sa povećanjem parametra N i u uslovima velikog šuma ($-10dB \le SNR \le 0dB$) dostiže svoj maksimum za N=8. Pri velikom SNR-u ($5dB \le SNR \le 20dB$) maksimum za N=8 nije evidentiran. Dalje povećanje parametra N dovodi do snižavanja $\overline{MSE}$.

f. Minimum $\overline{MSE}$ eksponencijalno opada sa porastom SNR-a. Najniža vrednost $\overline{MSE}$ dobija se za N=64 pri SNR=0dB. Za ovu vrednost parametra N, dalje povećanje SNR-a izaziva oscilovanje minimuma $\overline{MSE}$ oko vrednosti 0.084.

V. ZAKLJUČAK

U radu su određene performanse MFCC algoritma za rekonstrukciju govornog signala kome je superponiran akustički Babble šum. Detaljna analiza je pokazala da se minimum $\overline{MSE}$ dobija za dužinu MFCC vektora I=23. Takođe, povećanje SNR-a ili smanjenje broja govornika (N<8) dovodi do snižavanja $\overline{MSE}$. Za N=8 ima se maksimum $\overline{MSE}$. Povećanje broja govornika (N>8) dovodi do snižavanja $\overline{MSE}$. Tako, najveća vrednost $\overline{MSE}$ dobijena je za N=8 u a najniža za N=64 u testiranom SNR opsegu. U nastavku istraživanja ocena kvaliteta rekonstruisanih govornih signala izvršiće se MOS testom.

LITERATURA

[1] J.M. Huerta and R.M. Stern, "Speech recognition from GSM codec parameters", *Proceedings of ICSLP*, pp. 1463-1466, 1998.

[2] D. Chazan, G. Cohen, R. Hoory, M. Zibulski "Low Bit Rate Speech Compression for Playback in Speech Recognition Systems", *EUSPICO*, pp. 1281–1284, Sep. 2000.

[3] "*ES 201 108 – STQ : DSR – Front-end feature extraction algorithm ; compression algorithm*", ETSI document – 2000.

[4] B. Milner and X. Shao, "Clean speech reconstruction from MFCC vectors and fundamental frequency using an integrated front-end", *Speech Communication* 48 (2006) 697–715.

[5] M. Slaney, "Auditory toolbox version 2", Tech. Rep.1998-010, Interval Research Corporation, 1998.

[6] D. Chazan, G. Cohen, R. Hoory, and M. Zibulski, "Speech reconstruction from mel frequency cepstral coefficients and pitch", *Proceedings of ICASSP* 2000.

[7] T. Nakatani, S. Amano, T. Irino, K. Ishzuka, T. Kondo, "A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments", *Speech Communication* 50 (2008) pp. 203–214.

[8] J. Rouat, Y.C. Liu, D. Morissette, "A pitch determination and voiced/unvoiced algorithm for noisy speech", *Speech Commun. J.*, (1997), pp. 191–207.

[9] *"ES 202 212—STQ: DSR, Extended advanced front-end feature extraction algorithm; compression algorithms; back-end speech reconstruction algorithm"*, ETSI document, 2003.

[10] R.J. McAuley and T.F. Quatiery, "Speech analysis/synthesis based on a sinusoidal representation", *IEEE Transaction Acoust.,Speech, Signal Processing*, vol. 34, pp. 744-754, 1986.

[11] Z. Veličković, Z. Milivojević, "Performaces of the MFCC algorithm", *INFOTEH-JAHORINA* 2008, Vol. 7, pp. 180-184.

[12] P. Mokhtari, H. Takemoto, T. Kitamura, "Single-matrix formulation of a time domain acoustic model of the vocal tract with side branches", *Speech Communication* 50 (2008), pp. 179-190.

[13] D. Ellis, "Speech & Audio Processing & Recognition", Available: http://www.ee.columbia.edu/~dpwe/e6820/, 2006.

[14] Z.N. Milivojević, M.Dj. Mirković. "Estimation of the fundamental frequency of the speech signal modeled by the SYMPES method", *Int J Electron Commun (AEU)*, 2008, doi: 10.1016/j.aeue.2007.12.006.

[15] S. Simpsona, M. Cooke, "Consonant identification in N-talker babble is a nonmonotonic function of N (L)", *J. Acoust. Soc. Am.* 118 (2005), pp. 2775-2778.

ABSTRACT

In this paper, we determined the performances of a MFCC algorithm in Babble noise. The performances of the MFCC algorithm are determined by MSE of original and reconstructed speech in variable SNR and MFCC vector length. The sinusoidal model of speech reconstruction based on approximated amplitude characteristic and fundamental frame frequency is used. The MSE of additive Babble noise from one, two, four, eight and sixty four speakers are analyzed. The comparative analyze of the MSE results are used.

SPEECH RECONSTRUCTION FROM MFCC IN BABBLE NOISE

Zoran Veličković, Zoran Milivojević, Danijela Aleksić