Stereo akustična lokalizacija aktivnog govornika

Dragan D. Kukolj, *Senior Member, IEEE,* Ištvan I. Papp, Saša A. Vukosavljev, Vladimir R. Djurkovic

***Sadržaj* — Procena dolaznih pravaca (*DOA*, *Direction Of Arrival*) je oblast istraživanja od velikog praktičnog interesa (radari, sonari, "hands-free" govorna komunikacija, itd.). Ta oblast je doživela veliki razvoj u proteklih nekoliko dekada i nastavlja da se razvija. U ovom radu je izloženo jedno proširenje algoritma za procenu dolaznih pravaca koja se zasniva na metodi PHAT Generalizovane kros-korelacije (*Generalized Cross-correlation*) koja je posebno primenljiva za prostorije sa prisutnim odjekom i za ograničene računarske resurse, jer pri proceni koristi informacije sa samo dva mikrofona.**

***Ključne reči* — Procena dolaznih pravaca, generalizovana kros-korelacija, procena vremenskih kašnjenja, fazna transformacija.**

I. Uvod

Procena dolaznih pravaca, uz korišćenje prostorno razdvojenih mikrofona je povezana sa raznim primenama kao što su: video telefonski sistemi, telekonferencijski sistemi, "hands-free" govorna komunikacija [1]-[4], sistemi za poboljšanje kvaliteta i razumljivost govornog signala, robotika, prepoznavanje govora, sistemi za praćenje i nadzor. U svim tim primenama, procena dolaznih pravaca primenjuje se u cilju lokalizacije aktivnog i dominantnog govornika. U ovom radu, opisano je jedno poboljšanje metode generalizovanih kros-korelacija sa akcentom na primenu u prostorijama sa prisutnim većim odjekom i većim nivoom ambijentalnog šuma. Posebno je podesna za primenu u robotici i sličnim sistemima sa ograničenim resursima, jer pri proceni dolaznih pravaca koristi informaciju samo dva mikrofona.

Postoje tri tipa metoda za procenu dolaznih pravaca: (a) Maksimizacija izlazne snage izlaza iz superdirektivnog mikrofonskog niza po dolaznim pravcima [5], (b) Visokorezolutivna procena spektra, i (c) Procena vremenskih kašnjenja. Prvi pristup se retko koristi za procenu dolaznih pravaca zbog velike računske složenosti postupka. Drugi pristup, tj. visoko rezolutivna procena spektra koristi prostorno spektralnu korelacionu matricu koja se formira na osnovu mikrofonskih signala. Veliki broj sub-optimalnih tehnika sa redukovanom kompleksnošću iz ove klase je dobro poznat. Tu spadaju: metod minimalne varijanse, metod minimalne norme, metod višestruke klasifikacije signala (MUSIC) [6], itd.

Metode procene vremenskih kašnjenja [7] (*Time-delay Estimation, TDE*) zasnovan je na proceni vremenskih kašnjenja između mikrofonskih signala i u praksi se najviše koristi. Zasniva se na lokalizaciji maksimuma kros-korelacione funkcije između para mikrofonskih signala. Metoda generalizovane kros-korelacije [8] je poznata metoda iz klase TDE metoda za procenu dolaznih pravaca i zasniva se na lokalizaciji maksimuma kros-korelacione funkcije između ponderisanih mikrofonskih signala. Funkcija se ponderiše zbog povećanja robusnosti algoritma procene dolaznih pravaca na prisustvo šuma i odjeka u prostoriji. Dve najčešće korišćene težinske funkcije su: funkcija raspodele maksimalne verodostojnosti (*ML, Maximum Likelihood*) i funkcija fazne transformacije (*PHAT, Phase Transform*). Dok funkcija maksimalne verodostojnosti ističe signal na frekvencijama gde je odnos signala i šuma veliki (*SNR, Signal to Noise Ratio*), primenom funkcije fazne transformacije, poravnava se amplitudno-spektralna karakteristika signala [9], pri čemu se takođe sistem čini invarijantnim na snagu mikrofonskih signala, jer se kros-korelacija u spektru normira. Iako je praksa pokazala da se sa PHAT funkcijom postižu dobri rezultati samo kada je SNR dovoljno veliki, proširenje PHAT funkcije iskorišćeno je kao osnova za nadgradnju algoritma zbog ostalih dobrih osobina.

Rad je organizovan u četiri poglavlja. Drugo poglavlje predstavlja generalni prikaz algoritma i njegovih modula. U trećem poglavlju detaljno je opisan princip rada pojedinih algoritamskih modula, dok četvrto poglavlje opisuje performanse kao i proceduru verifikacije algoritma. Peto poglavlje je zaključak.

II. Pregled Kompletnog Algoritma

Osim glavnog algoritamskog modula, čiju osnovu čini proračun PHAT težinske funkcije ,za robusnu procenu DOA, opisani algoritam ima modul za detekciju govorne aktivnosti (*VAD, Voice Activity Detector*), modul za favorizovanje frekvencija sa većim odnosom signal šum (*NM, Noise Masking*), modul za ocenu zvučnosti govornih segmenata (*SFE*, *Spectral Flatness Estimator*), kao i algoritam za klasifikaciju, tj. klasterovanje trenutnih vrednosti DOA (*CL, Clustering Algorithm*).

Glavni modul algoritma za lokalizaciju aktivnog govornika je PHAT algoritamski blok koji predstavlja modul za direktnu procenu DOA, i spada u klasu algoritama koji procenjuju dolazni pravac preko procene vremenskih kašnjenja (*TOA, Time Of Arrival*) između

Rad je delimično podržan u okviru projekta TR-6136B od strane Ministarstva za nauku i zaštitu životne sredine Republike Srbije.

D. D. Kukolj, Fakultet tehničkih nauka u Novom Sadu, Srbija (telefon: 381-21-4801141; e-mail: dragan.kukolj@micronasnit.com).

I. I. Papp Fakultet tehničkih nauka u Novom Sadu, Srbija (e-mail: istvan.papp@micronasnit.com).

S. A. Vukosavljev, Micronas NIIT, Fruškogorska 11, Novi Sad, Srbija (e-mail: sasa.vukosavljev@micronasnit.com).

Vladimir. R. Durkovic, Micronas NIIT, Fruškogorska 11, Novi Sad, Srbija (e-mail: vladimir.djurkovic@micronasnit.com).

mikrofonskih signala. Pošto je algoritam namenjen za sistem sa dva mikrofona, PHAT procenjuje razliku u vremenu pristizanja izvornog signala koji potiče od aktivnog govornika na prvi i drugi mikrofon i na osnovu te procene i rastojanja mikrofona daje procenu dolaznog ugla signala izvora koji direktno predstavlja dolazni pravac aktivnog govornika. Procenu dolaznih pravaca PHAT algoritamski modul vrši za svaki pojedinačni segment mikrofonskih signala u vremenu, pri čemu se dobija vremenska serija procenjenih dolaznih pravaca aktivnog govornika.

Svi ostali navedeni algoritamski moduli, osim PHAT modula, imaju za cilj da se poveća robusnost algoritma na prisustvo šuma i odjeka u prostoriji i kombinacijom kriterijuma koji oni pojedinačno definišu, dobija se generalni kriterijum, koji određuje validni DOA iz vremenske serije trenutno procenjenih DOA izlaza PHAT modula.

VAD modul procenjuje srednju snagu mikrofonskih signala na mikrofonima i ako je srednja snaga iznad određenog praga aktivira PHAT algoritamski modul da za dati vremenski blok mikrofonskih signala proceni dolazni pravac aktivnog govornika.

NM algoritamski modul formira težinsku funkciju koja favorizuje DFT koeficijente mikrofonskih signala gde je najpovoljniji odnos signala i šuma. Za formiranje težinske funkcije za tekući *k*-ti vremenski blok, koristi se procena šuma za dati blok, koja se formira na osnovu srednje vrednosti mikrofonskih signala za prethodni *(k-1)*-vi blok.

SFE algoritamski blok koristi se u cilju povećanja robusnosti algoritma na šum u prostoriji. SFE blok procenjuje zvučnost pojedinih segmenata mikrofonskih signala u cilju odredjivanja zvučnih glasova u govornom signalu koji dopire od aktivnog govornika. Samo segmenti u vremenu sa zadovoljenim kriterijumom zvučnosti se uzimaju u obzir pri određivanju validnog DOA.

CL blok za klasifikaciju se koristi za isključivanje trenutnih vrednosti DOA sa malom verovatnoćom pojave, koja je procenjena na osnovu učestanosti pojavljivanja.

III. PRINCIP RADA POJEDINIH ALGORITAMSKIH MODULA

Glavni algoritamski modul opisanog algoritma procene dolaznih pravaca je PHAT algoritam za procenu vremenskih kašnjenja, dok ostali algoritamski moduli povećavaju robusnost algoritma na prisutni šum i odjek.

A. Detektor govorne aktivnosti (VAD)

Detektor govorne aktivnosti radi na osnovi jednostavnog algoritma određivanja energije kratkotrajnih vremenskih segmenata, koje je realizovano u spektru. Pošto je na osnovu Parseval-ove teoreme za diskretne signale energija *k*-tog vremenskog segmenta diskretnog signala *E* jednaka energiji njegovog spektra, onda u slučaju da za *k*-ti vremenski segment važi $E > T_h$ (T_h - unapred definisani prag energije koji je određen heuristički), VAD algoritamski blok indicira da je otkrivena govorna aktivnost.

B. Fazna transformacija (PHAT)

Fazna transformacija (*PHAT*) je naziv za jednu od varijanti metode generalizovane kros-korelacije (*GCC*) i najkorišćenija je metoda za procenu vremenskih kašnjenja (*TDE*) [5], prvenstveno zbog svoje jednostavnosti, ali i veće robusnosti na odjek u prostoriji (uz korišćenje odgovarajuće težinske funkcije, kao što je *PHAT*) u odnosu na većinu složenijih postupaka. GCC metoda procenjuje kašnjenje po sledećoj relaciji:

$$\hat{\tau}_{GCC} = \arg\max_n \hat{\psi}_{GCC}[n] \tag{1}$$

gde je:

$$\hat{\psi}_{GCC}[n] = \sum_{k=0}^{N-1} \phi[k] S_{x_0 x_1}[k] e^{j\frac{2\pi nk}{N}} , \tag{2}$$

pri čemu je *N* dužina bloka, $S_{x_0x_1}[k] = E\{X_0[k]X^*_1[k]\}$ je kros-spektar mikrofonskih signala, a operatori $E\{.\},(.)^*$ predstavljaju matematičko očekivanje i kompleksno konjugovanje, respektivno. Oznaka $X_l[k], l \in \{0,1\}$ predstavlja DFT signala prvog i drugog mikrofona, dok $\phi[k]$ predstavlja spektralnu težinsku funkciju po kojoj se varijante GCC metode i razlikuju. Iz priloženih relacija se vidi da $\hat{\psi}_{GCC}[n]$ predstavlja procenu kros-korelacione funkcije između mikrofonskih signala, tako da se traženjem argumenta te funkcije koji je maksimizira, dobija procena kašnjenja.

Generalno, kros-spektar stacionarnih slučajnih serija se može predstaviti kao Furijeova transformacija kros-korelacione funkcije $S_{x_0x_1}(k) = \Im\{r_{x_0x_1}(n)\}$, pri čemu predstava $S_{x_0x_1}[k] = X_0[k]X^*_1[k]$ važi za determinističke signale, dok za slučajne serije oblik $S_{x_0x_1}[k] = E\{X_0[k]X^*_1[k]\}$ važi samo aproksimativno. Tako se gornjom relacijom (2), sa $\hat{\psi}_{GCC}[n]$ samo aproksimira kros-korelaciona funkcija $r_{x_0x_1}(n)$ po svojoj prirodi slučajnih mikrofonskih signala x_0 i x_1. Izraz $S_{x_0x_1}[k]$ se procenjuje rekurzivno sa eksponencijalnim prozorom zaboravljanja.

Težinska funkcija ima bitnog udela na performanse TDE procene primenom GCC algoritma. Osim jedinične težinske funkcije, u čestoj upotrebi su i ML težinska funkcija kojom se dobija efikasna procena TDE, tj. varijansa procene dostiže granicu Rao-Cramer kao donju granicu, kao i PHAT težinska funkcija. PHAT težinska funkcija je definisana kao:

$$\phi_{PHAT}[k] = \frac{1}{\left|S_{x_0x_1}[k]\right|} . \tag{3}$$

Vidi se da se primenom PHAT težinske funkcije normalizuje spektar po modulu, pri čemu onda u filtriranom kros-spektru figurišu samo fazna kašnjenja mikrofonskih signala, čime je procena invarijantna na snagu signala.

C. Maskiranje šuma

Činjenica je da u proceni TDE metodom kros-spektralne korelacije, na rezultujuću korelaciju sve frekvencije utiču podjednako, čak i ako je na datoj frekvenciji dominantan šum. To čini algoritam manje robusnim na šum i otežava

procenu vremenskih kašnjenja.

U cilju da se prevaziđe taj problem, primenjena je spektralna težinska funkcija $w[k]$. Ova funkcija daje veći značaj frekvencijama sa većom vrednosti SNR-a. Sum se maskira za *k*-tu diskretnu frekvenciju na sledeći način:

$$w[k] = \max\left\{0.1, \frac{X[k] - \alpha X_n[k]}{X[k]}\right\}, \quad 0 < \alpha < 1 \tag{4}$$

gde je $X[k]$ srednja vrednost spektralne gustine snage, a $X_n[k]$ procena šuma za *k*-tu diskretnu frekvenciju. Koeficijenti $w[k]$ su bliski jedinici na frekvencijama gde je korisni signal znatno jači od šuma. U cilju pojačavanja ovog efekta težinska funkcija se modifikuje sa:

$$w_e[k] = w[k]\left(\frac{X[k]}{X_n[k]}\right)^{\gamma} \tag{5}$$

gde $0 < \gamma < 1$ daje veću težinu frekvencijama sa velikom vrednosti SNR mere.

D. *Ocena zvučnosti govornih segmenata*

U cilju povećanja pouzdanosti lokalizacije aktivnog govornika, pored VAD modula zasnovanog na merenju snage signala, trenutno procenjeni TDE se smatra validnim samo u periodima kada su prisutni zvučni glasovi. Modul za prepoznavanje zvučnih glasova koristi ocenu nivoa ujednačenosti spektra (*SF*, *spectral flatness*), koja predstavlja meru zašumljenosti, dekorelisanosti i ujednačenosti spektra ili jednog njegovog dela. Računa se kao odnos između geometrijske i aritmetičke sredine energije spektra signala, odnosno,

$$SF = \frac{\left(\prod_{i=1}^{N} A(i)\right)^{1/N}}{\frac{1}{N}\sum_{i=1}^{N} A(i)} \tag{6}$$

gde je *A(i)* amplituda *i*-tog frekventnog opsega. Vrednosti *SF* mere su manje za zvučne glasove.

E. *Algoritam za klasifikaciju*

Klasifikacija je grupisanje uzoraka *(Pattern clustering)* u određeni broj homogenih grupa (klasa) na bazi odabrane mere sličnosti između uzoraka. Uzorci klasifikovani u istu klasu treba da budu slični jedni drugima, dok to ne treba da važi za uzorke iz različitih klasa. U slučaju procene kašnjenja, klasifikacija je jednodimenzionalni problem nad vremenskom serijom. Svaka tačka te serije predstavlja vremensko kašnjenje TDE kao rezultat PHAT korelacije. Algoritam klasifikacije generiše stabilne i pouzdane procene vremenskih kašnjenja kroz sledeće korake:

(1) Algoritam klasifikacije startuje kada ulazni bafer D sadrži N poslednjih validnih rezultata procene PHAT korelacijom. Ako je bafer već pun, novi TDE iz PHAT estimatora zamenjuje najstariji. Broj postojećih klasa je postavljen na G = 1 i centar te klase je postavljen na srednju vrednost ulaznog bafera. Za svaki novi, *k*-ti TDE uzorak, obavljaju se naredni koraci:

(2) Računa se srednja vrednost *d* i varijansa *s* poslednjih *N* vrednosti iz ulaznog bafera ***D*** kao:

$$d = \tfrac{1}{N}\sum_{i=1}^{N} D(i) \quad , \quad s = \tfrac{1}{N}\sum_{i=1}^{N} (D(i) - d)^2 \tag{7}$$

Taj korak igra ulogu usrednjavanja ulaznih uzoraka.

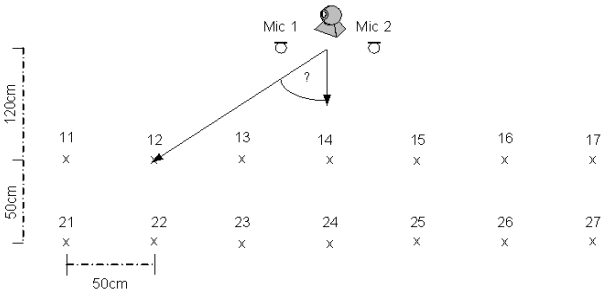
(3) Ako je vrednost varijanse *s* ispod prethodno definisanog praga *T1*, trenutna srednja vrednost *d* elemenata bafera ***D*** se dodeljuje najbližoj postojećoj klasi ***m***(*w*), tj. $d \in m(w)$, ako važi $\|d - m(w)\| < \|d - m(g)\|$, *g*=1,..,*G*, i $\|d - m(w)\| < T_2$, gde je $\|.\|$ Euklidska norma u ***R*** i gde je prag *T2* izabran tako da nema preklapanja između klasa. Ako se sa ***p*** predstavi vektor koji indicira kada je poslednji put neki TDE podatak uvršten u koju od klasa, onda komponenta vektora ***p*** za dato *g* (*g* je indeks klase koja je primila *d*) postaje ***p***(*g*) = *k*.

(4) Ukoliko je ipak vrednost varijanse *s* iznad praga *T1*, ili nije ispunjen uslov najbliže klase $\|d - m(w)\| < T_2$, tada se formira nova klasa sa centrom $m(G) = d$ i izvrši se operacija *G* = *G*+1, $G \le G_{max}$, gde je G_{max} maksimalno dozvoljen broj klasa. Takođe, ako je ispunjen uslov $(G > G_{\max}) \vee (\exists g, (k - p(g)) > T_{\max})$, tada se "najstarija" klasa uklanja iz skupa klasa. $T_{\max}$ je najveći dozvoljeni broj iteracija koji može da prođe od kada je nekoj klasi dodeljen neki podatak *d*, i to važi za sve klase.

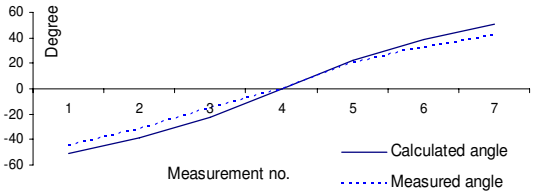
Koraci od (2)-(4) se ponavljaju za svaki *k*-ti TDE uzorak dobijen kao validni rezultat PHAT korelacije.

IV. VERIFIKACIJA ALGORITMA

Da bi se verifikovao algoritam realizovana je hardverska platforma i izvedene su dve grupe eksperimenata. Hardverska platforma je između ostalog sadržala: PC računar, zvučnu karticu, dva mikrofona na rastojanju od 25cm i kameru. Signal je uzorkovan na 48 kHz. Za potrebe prve grupe eksperimenata je formirana ekvidistantna mreža tačaka na podu prostorije, kao što je šematski prikazano na Sl. 1. Aktivni govornik se kretao po čvorovima mreže i pri tome su rezultati procene DOA (u uglovima u stepenima) dobijeni sa implementiranim algoritmom, poređeni sa stvarnim vrednostima tih uglova. Rezultati eksperimenta su u obliku apsolutne greške procene prikazani na Sl. 2.



Sl. 1. Eksperiment 1: Horizontalni prikaz položaja govornik - sistem za lokalizaciju.

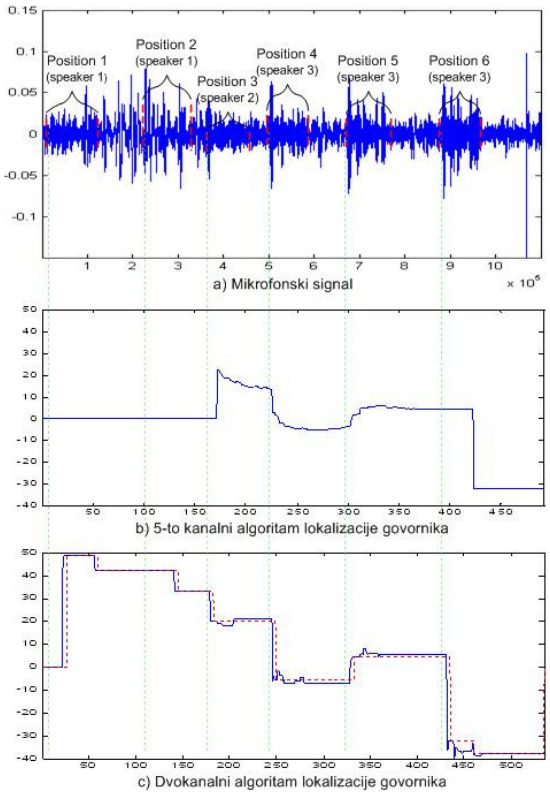


Sl. 2. Preciznost lokalizacije govornika kod prvog eksperimenta.

U drugoj grupi eksperimenata je poređen opisani pristup sa algoritmom za procenu DOA zasnovanim na mikrofonskom nizu. Korišćena je konfiguracija sistema sa 5 mikrofona na rastojanju od 5 cm i sa frekvencijom odabiranja signala od 11,025kHz. Ovaj algoritam se takođe zasniva na PHAT korelaciji, ali se korelacija primenjuje na parove mikrofonskih signala (1,2), (1,3) ,..., (1,5), pri čemu su korelacione funkcije pojedinih mikrofonskih parova pre sabiranja interpolirane u vremenu, da bi se uskladile po kašnjenjima. Ovo algoritamsko rešenje takođe sadrži i superdirektivni mikrofonski niz (*beamformer*) koga čine tih 5 mikrofona. U tom algoritmu se validnost trenutnog TDE određuje na osnovu odnosa snage signala iz superdirektivnog mikrofonskog niza za trenutni DOA i srednje snage mikrofonskih signala. Ako je taj odnos povoljan trenutni DOA se proglašava validnim.

TABELA 1: POZICIJE GOVORNIKA

Pozicija	Govornik	Koordinate u prostoriji [cm]	Ugao [stepeni]
1	1	(210,140)	56
2	1	(210,250)	40
3	2	(90, 310)	16
4	3	(-40,260)	-9
5	3	(0, 140)	0
6	3	(-100,90)	-48



Sl. 3. Eksperiment 2: Primer lokalizacije govornika sa promenom pozicije.

Eksperiment za poređenje dva algoritma je obavljen na sledeći način: U prostoriji zapremine 140m^3 su tri govornika govorila sa više različitih pozicija (ukupno šest) i sa promenljivim SNR u opsegu od 4÷5dB. Položaj govornika u odnosu na centralnu osu mikrofonskog niza sistema su dati u Tabeli 1. Položaji govornika su procenjeni sa oba algoritma. Rezultat poređenja je dat na Sl. 3. Kod dvokanalnog algoritma je kašnjenje u proseku oko 0.2 s, a kod 5-kanalnog je nešto veće. U ovom slučaju 5-kanalni algoritam nije odredio obe pozicije prvog govornika (tiši govornik), dok je tačnost i ponašanje procene ostalih govornika kod oba algoritma zadovoljavajuća i pokazuje male varijanse.

V. ZAKLJUCAK

Rad predstavlja jedno proširenje algoritma za procenu dolaznih pravaca zasnovanoj na metodi PHAT generalizovane kros-korelacije. Pri proceni dolazanih pravaca koriste se informacije sa samo dva mikrofona, te je pomenuta metoda posebno primenljiva za sisteme sa ugradjenom aplikacijom i ograničenim računarskim resursima.

LITERATURA

[1] J.-M. Valin, F. Michaud, J. Rouat, D. Létourneau, Robust sound source localization using a microphone array on a mobile robot, *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, 2003.

[2] R. L. B. Jeannμes, P. Scalart, G. Faucon, and C. Beaugeant, Combined noise and echo reduction in hands-free systems: A survey, *IEEE Trans. Speech Audio Proc.*, Vol.9,pp.808–820, 2001.

[3] I. Pap, D. Kukolj, Z. Marčeta, V. Đurkovic, M. Janev, M. Popović, N. Teslić, Remotely controlled semi-autonomous robot with multimedia abilities, *ICCA 2005*, Budapest, June 26-29, 2005.

[4] E. Mumolo, M. Nolich, G. Vercelli, Algorithms for acoustic localization based on microphone array in service robotics, *Robotics and Autonomous Systems,* Vol. 42, pp. 69–88, 2003.

[5] J. Benesty and Y. Huang, *Audio signal processing for next generation multimedia communication systems*. Boston: Cluwer Academic Publ., 2004.

[6] S. L. Marple, Jr., D*igital spectral analysis with applications*. NJ: Prentice-Hall, 1987.

[7] P. Julian et al., A comparative study of sound localization algorithms for energy aware sensor network nodes, *IEEE Trans. Circuits and Systems*, Vol. 51, No. 4, pp. 640-648, Apr. 2004.

[8] C. H. Knapp and G. C. Carter, The generalized correlation method for estimation of time delay, *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-24, pp. 320–327, Aug. 1976.

[9] B. Mungamuru and P. Aarabi, Enhanced sound localization, *IEEE Trans. Syst., Man, Cybern.—Part B:Cybern.*, Vol.34, No.3, 2004.

ABSTRACT

***Abstract* — Direction of arrival estimation has been an area of intensive research and practical interest for many decades (radar, sonar, hands-free communication, etc.), and it is in continuous progress. In this paper, one extension of the algorithm based on PHAT Generalized Cross-Correlation is presented. The presented method is especially applicable in noisy and reverberant conditions, in systems with limited computational resources, since the estimation is using only two microphones.**

STEREO SOUND LOCALIZATION OF THE ACTIVE SPEAKER

Dragan D. Kukolj, *Senior Member, IEEE,* Ištvan I. Papp, Saša A. Vukosavljev and Vladimir R. Djurkovic