Formantinių požymių naudojimas kalbai atpažinti

Antanas Leonas Lipeika

Matematikos ir informatikos instituto
Atpažinimo procesų skyriaus vyresnysis mokslo darbuotojas, profesorius, daktaras
Institute of mathematics and informatics, Senior Researcher, Prof., PhD
Goštauto g.12, LT-01108 Vilnius
Tel.(8 5) 266 0390, faks. (8 5) 261 9905
El. paštas: lipeika@ktl.mii.lt

Straipsnyje nagrinėjami formantinių požymių taikymo atpažįstant kalbą klausimai. Nustatyta, kad formantiniai požymiai tam gali būti naudojami, tačiau atpažinimo tikslumas labai priklauso nuo formantinių požymių išskyrimo metodo. Geriausi atpažinimo rezultatai gaunami formantinių požymių išskyrimui naudojant išsigimusius prognozės polinomus. Šie polinomai gali būti skaičiuojami iš lyginės arba nelyginės eilės tiesinės prognozės modelio parametrų. Be to, atpažinimui galima naudoti simetrinius arba antisimetrinius išsigimusius tiesinės prognozės polinomus. Taip pat svarbu ištirti, kaip kalbos atpažinimo rezultatai priklauso ne tik nuo išsigimusių tiesinės prognozės polinomų parinkimo, bet ir kitų atpažinimo sistemos parametrų: analizės kadro ilgio, atpažinimui naudojamų formančių skaičiaus, formantinių požymių vaizdavimui naudojamos dažnių skalės. Tyrimais nustatyta, kad geriausi atpažinimo rezultatai gaunami naudojant dvi arba tris formantes, apskaičiuotas iš simetrinių išsigimusių prognozės polinomų. Tiriant atskirų formančių informatyvumą paaiškėjo, kad didžiausias indėlis į atpažinimą yra antros formantės. Pirmos, trečios ir ketvirtos formančių indėlis maždaug vienodas, bet aukštesnės formantės mažiau atsparios balto triukšmo įtakai. Tiriant analizės kadro ilgio parinkimą nustatyta, kad geriausi atpažinimo rezultatai yra esant 500 atskaitų kadro ilgiui. Atpažinimo rezultatai taip pat gaunami geresni vaizduojant formančių trajektorijas melų skalėje.

Pripažįstama, kad formantiniai požymiai gali būti naudojami kalbai atpažinti, tačiau jie iki šiol nebuvo populiarūs dėl jų išskyrimo problemų. Mėginta formantinius požymius naudoti paslėptais Markovo modeliais grįstame kalbos atpažinime (De Wet et al., 2004), bet buvo pasiektas tik nereikšmingas atpažinimo tikslumo padidėjimas. Mes bandėme formantinius požymius taikyti dinaminiu laiko skalės kraipymu grįstam izoliuotų žodžių atpažinimui (Lipeika, 2005). Tiriant išryškėjo formantinių požymių išskyrimo patikimumo problema. Formančių trajektorijos ne visada būdavo glotnios, artimos formantės kartais susiliedavo ir iškildavo jų numeravimo problemų. Kadangi formančių dažnių įverčiai skaičiuojami iš tiesinės prognozės modelio parametrų, ieškojome patikimesnio šio modelio parametrų vertinimo metodo (Lipeika, 2007). Įprastai naudojamas autokoreliacinis tiesinės prognozės modelio parametrų vertinimo metodas buvo lyginamas su kovariaciniu, Burgo, Marple'o metodais ir modifikuotu Split Levinsono algoritmu. Tiriant nustatyta, kad formančių trajektorijų išskyrimo požiūriu kovariacinis, Burgo, Marple'o tiesinės prognozės modelio parametrų vertinimo metodai iš esmės nesiskiria nuo autokoreliacinio, o modifikuotu Split Levinsono algoritmu gauname daug patikimesnius formančių trajektorijų įverčius. Formančių trajektorijų įverčiai gaunami glotnesni ir algoritmas garantuoja, kad išskirtų formančių skaičius visada bus lygus pusei tiesinės prognozės mode-

lio eilės. Tai padeda atpažinimo metu išvengti formančių numeravimo klaidų.

Split Levinsono algoritmas remiasi vadinamųjų išsigimusių tiesinės prognozės polinomų skaičiavimu (Delsaite, Genin, 1986). Jeigu z transformacijos srityje turime tiesinės prognozės polinomų aibę

$$A_k(z) = 1 + a_k(1)z^{-1} + a_k(2)z^{-2} + \cdots$$
$$\cdots + a_k(k)z^{-k}, k = 1, \cdots, p \quad , \qquad (1)$$

polinomai yra susieti priklausomybe

$$A_{k+1}(z) = A_k(z) + \rho_{k+1} z^{-(k+1)} A_k(z^{-1}), \qquad (2)$$

čia $\rho_1, \rho_2, \ldots, \rho_p$ yra atspindžio koeficientai; $a_k(1)$, $a_k(2)$ …, $a_k(k)$ yra k-osios eilės tiesinės prognozės polinomo koeficientai; z tolydinis kompleksinis kintamasis. Jeigu atspindžio koeficientą ρ_{k+1} prilyginsime 1 arba -1, tai yra ekvivalentiška prielaidai, kad atitinkamas balso trakto akustinio vamzdžio modelis $p+1$-oje pakopoje yra visiškai uždarytas ($\rho_{k+1} = 1$) arba visiškai atidarytas ($\rho_{k+1} = -1$) ir iš (2) gausime du išsigimusius prognozės polinomus:

$$P_{k+1}(z) = A_k(z) + z^{-(k+1)} A_k(z^{-1}) =$$
$$= 1 + (a_k(1) + a_k(k))\, z^{-1} + (a_k(2) + a_k(k-1))\, z^{-2} + \cdots$$
$$\cdots + (a_k(k) + a_k(1))\, z^{-k} + z^{-(k+1)} \qquad (3)$$

ir

$$Q_{k+1}(z) = A_k(z) - z^{-(k+1)} A_k(z^{-1}) =$$
$$= 1 + (a_k(1) - a_k(k))\, z^{-1} + (a_k(2) - a_k(k-1))\, z^{-2} + \cdots$$
$$\cdots + (a_k(k) - a_k(1))\, z^{-k} - z^{-(k+1)}. \qquad (4)$$

Polinomas $P_{k+1}(z)$ yra simetrinis, $Q_{k+1}(z)$ – antisimetrinis. Todėl

$$A_k(z) = 1/2[P_{k+1}(z) + Q_{k+1}(z)]. \qquad (5)$$

Naudodami (3), (4) išraiškas iš p-os eilės tiesinės prognozės polinomo galime gauti du $p+1$ eilės išsigimusius prognozės polinomus $P_{p+1}(z)$ ir $Q_{p+1}(z)$. Šių polinomų šaknys yra ant vienetinio apskritimo, todėl tiesinės prognozės modelio amplitudiniame spektre, apskaičiuotame iš šių polinomų, ties dažniais, lygiais šaknų kampui su realia ašimi, atsiranda smailūs pikai. Šie dažniai yra traktuojami kaip formančių dažnių įverčiai ir yra naudojami kaip kalbos atpažinimo požymiai (Lipeika, Lipeikienė, 2008).

Spektro pikų vietos priklauso nuo to, ar išsigimusiems prognozės polinomams skaičiuoti naudojame lyginės, ar nelyginės eilės tiesinės prognozės modelį. Be to, spektro pikų vietos priklauso nuo to, ar mes naudojame simetrinį išsigimusį prognozės polinomą, ar antisimetrinį (Kabal, Ramachandran, 1986). Todėl reikalingas tyrimas nustatyti, kokį išsigimusį prognozės polinomą naudoti ir kokią parinkti išsigimusio prognozės polinomo eilę, taip pat ištirti atskirų formančių informatyvumą, kiek formančių naudoti atpažinimui, kaip kadro ilgis veikia atpažinimo tikslumą, kaip atpažinimo tikslumas priklauso nuo dažnių skalės.

Formantinių požymių taikymo žodžių atpažinimui eksperimentinis tyrimas

Tyrime buvo naudojama 111 lietuvių kalbos žodžių: 99 žodžiai buvo atrinkti iš dažninio dabartinės rašomosios lietuvių kalbos žodyno (Grumadienė, Žilinskienė, 1997), kiti žodžiai buvo skaičiai nuo nulio iki devynių ir žodžiai „pradžia“, „pabaiga“. Garsai buvo įrašyti įprastinėje kambario aplinkoje, esant 30 dB signalo ir triukšmo santykiui. Etaloniniai žodžių garso įrašai nebuvo papildomai užtriukšminami. Kad ištirtume atpažinimo sistemos atsparumą triukšmui, prie testinių garso įrašų buvo pridedamas 65 dB, 60 dB arba 55 dB baltas triukšmas. Atpažinimui buvo naudojamas dinaminiu laiko skalės kraipymu grįstas metodas (Tamulevičius, Lipeika, 2003).

Kad išsiaiškintume, kiek formančių geriausia naudoti atpažinimui, buvo atliktas žodžių atpažinimo klaidų priklausomybės nuo atpažinimui naudojamų formančių skaičiaus tyrimas. Buvo parinkta 8-a tiesinės prognozės polinomo eilė (9-a išsigimusio prognozės polinomo eilė) ir simetriniam bei antisimetriniam išsigimusiems prognozės polinomams tiriama atpažinimo tikslumo priklausomybė nuo formančių skaičiaus.

Tyrimui buvo naudojami neužtriukšinti kalbos signalai ir signalai, prie kurių buvo pridėtas 65 dB triukšmas. Rezultatai vaizduojami 1 lentelėje.

1 lentelė. ***Atpažinimo klaidų skaičiaus (%) priklausomybė nuo formančių skaičiaus naudojant 9-os eilės simetrinius bei antisimetrinius išsigimusius prognozės polinomus***

Polinomas, formančių skaičius	**SNR=30 dB**	**+ 65 dB triukšmas**
Simetrinis, 2 formantės	0	13,5
Simetrinis, 3 formantės	0,9	10,8
Simetrinis, 4 formantės	1,8	28,8
Antisimetrinis, 2 formantės	3,6	33,3
Antisimetrinis, 3 formantės	1,8	32,4
Antisimetrinis, 4 formantės	4,5	64,8

Analogiškas tyrimas buvo atliktas 9-ai tiesinės prognozės polinomo eilei (10-a išsigimusio prognozės polinomo eilė) ir simetriniam bei antisimetriniam išsigimusiems prognozės polinomams. Tyrimo rezultatai vaizduojami 2 lentelėje.

2 lentelė. ***Atpažinimo klaidų skaičiaus (%) priklausomybė nuo formančių skaičiaus naudojant 10-os eilės simetrinius bei antisimetrinius išsigimusius prognozės polinomus***

Polinomas, formančių skaičius	**SNR=30 dB**	**+ 65 dB triukšmas**
Simetrinis, 2 formantės	2,7	11,7
Simetrinis, 3 formantės	0	11,2
Simetrinis, 4 formantės	0	19,8
Antisimetrinis, 2 formantės	1,8	25,2
Antisimetrinis, 3 formantės	0	19,8
Antisimetrinis, 4 formantės	0,9	42,3

Iš 1 lentelės matome, kad naudodami 9-os eilės išsigimusius prognozės polinomus mažiausiai atpažinimo klaidų gauname imdami pirmas dvi arba tris formantes, apskaičiuotas iš simetrinių išsigimusių prognozės polinomų. Tai galioja tiek papildomai neužtriukšmintam signalui, tiek pridėjus 65 dB triukšmą. Naudodami 10 eilės išsigimusius prognozės polinomus (2 lentelė) mažiausiai atpažinimo klaidų gauname imdami pirmas tris formantes, apskaičiuotas iš simetrinių išsigimusių prognozės polinomų.

Kad išsiaiškintume atskirų formančių indėlį į žodžių atpažinimą, 9-ai tiesinės prognozės polinomo eilei (10-a išsigimusio prognozės polinomo eilė) ir simetriniams išsigimusiems prognozės polinomams buvo tiriamas atskirų formančių informatyvumas pagal atpažinimo tikslumą. Tyrimo rezultatai vaizduojami 3 lentelėje.

3 lentelė. ***Atpažinimo klaidų skaičiaus (%) priklausomybė nuo formantės numerio naudojant 10-os eilės simetrinius išsigimusius prognozės polinomus***

Formantės eilės numeris	**SNR=30 dB**	**+ 65 dB triukšmas**
Pirma formantė	19,8	30,6
Antra formantė	1,8	28,8
Trečia formantė	17,1	63,9
Ketvirta formantė	17,1	87,3
Penkta formantė	51,3	97,3

Iš 3 lentelės matome, kad didžiausias indėlis į atpažinimą yra antros formantės. Neužtriukšmintam kalbos signalui naudodami vien antrą formantę, gauname tik 1,8 % klaidų. Pirmos, trečios ir ketvirtos formančių indėlis maždaug vienodas. Tačiau užtriukšmintam signalui geriausiai tinka pirma ir antra formantė kaip atspariausios triukšmui.

Žodžių atpažinimo sistemoje (Tamulevičius, Lipeika, 2003), naudojant tiesinės prognozės modelio arba kepstrinius požymius, tradiciškai buvo imamas 250 kalbos signalo atskaitų analizės kadras ir 11,025 kilohercų signalo diskretizavimo dažnis. Naudodami šį kadro ilgį formantiniais požymiais grįstam atpažinimui pastebėjome, kad formančių trajektorijos nėra pakankamai glotnios, ypač aukštesnių formančių. Todėl atlikome atpažinimo tikslumo priklausomybės nuo analizės kadro ilgio tyrimą. Formantiniams požymiams išskirti naudojome 10-os eilės simetrinius išsigimusius prognozės polinomus. Atpažinimui buvo naudojamos pirmosios trys formantės. Tyrimo rezultatai vaizduojami 4 lentelėje.

4 lentelė. ***Atpažinimo klaidų skaičiaus (%) priklausomybė nuo kadro ilgio naudojant 10-os eilės simetrinius išsigimusius prognozės polinomus***

Kadro ilgis	SNR= 30 dB	+ 65 dB triukšmas	+ 60 dB triukšmas	+ 55 dB triukšmas
N=250	0	11,2	29,7	63,9
N=400	0	10,8	27,0	55,8
N=500	0	8,1	27,9	51,3
N=600	0,9	9,0	27,0	51,3

Iš 4 lentelės matome, kad 250 atskaitų kadro ilgio tikrai nepakanka. Geriausius rezultatus gavome naudodami 500 atskaitų kadro ilgį.

Tyrimų metu pastebėjome, kad aukštesnių formančių trajektorijos yra labiau išsibarsčiusios negu žemesnių. Todėl nutarėme pabandyti formantes vaizduoti žmogaus klausai artimesnėje dažnių suvokimo skalėje – melų skalėje (Furui, 2001). Melų skalė yra artima logaritminei skalei dažniams, viršijantiems vieną kilohercą, ir artima tiesinei skalei dažniams iki vieno kiloherco. Melų skalė paprastai aproksimuojama naudojant išraišką

$$F_{mel} = 3321 \log_{10}(1 + F_{kHz}); \qquad (6)$$

čia F_{kHz} yra dažnis, išreikštas kilohercais.

Mes tyrėme žodžių atpažinimo klaidų priklausomybę nuo triukšmo lygio esant skirtingiems analizės kadro ilgiams, formančių trajektorijas vaizduodami melų skalėje. Tyrimui naudojome 9-os ir 10-os eilės simetrinius išsigimusius prognozės polinomus. Atpažinimui ėmėme pirmąsias tris formantes. Tyrimų rezultatai vaizduojami 5 lentelėje.

Jeigu palygintume rezultatus su atpažinimo rezultatais, gautais naudojant tiesinę dažnių skalę, tai tiek kadro ilgiui N=250, tiek N=500 melų skalėje atpažinimo rezultatus gavome geresnius. Lygindami tik melų skalėje gautus rezultatus matome, kad mažiausias atpažinimo klaidų skaičius gautas naudojant 10-os eilės simetrinius išsigimusius prognozės polinomus, esant 500 atskaitų kadro ilgiui.

Išvados

Atlikę formantinių požymių naudojimo atpažįstant kalbą tyrimus nustatėme, kad geriausi atpažinimo rezultatai gaunami imant dvi arba tris formantes, apskaičiuotas iš simetrinių išsigimusių prognozės polinomų. Tiriant atskirų formančių informatyvumą nustatyta, kad didžiausias indėlis į atpažinimą priklauso antrai formantei. Pirmos, trečios ir ketvirtos formančių indėlis maždaug vienodas. Tiriant analizės kadro ilgio parinkimą išaiškėjo, kad geriausi atpažinimo rezultatai yra esant 500 atskaitų kadro ilgiui. Atpažinimo rezultatai taip pat gaunami geresni vaizduojant formančių trajektorijas melų skalėje. Atlikdami tyrimus pastebėjome, kad žodžių atpažinimo tikslumas labai sumažėja užtriukšminus kalbos signalą pridėtiniu baltu triukšmu, todėl tolesnių tyrimų kryptis turėtų būti balto triukšmo poveikio atpažinimo tikslumui mažinimas.

5 lentelė. ***Atpažinimo klaidų skaičius (%) vaizduojant formantes melų skalėje ir naudojant 9-os ir 10-os eilės simetrinius išsigimusius prognozės polinomus. Kadro ilgis N=250 ir N=500. Išsigimusio prognozės polinomo eilė M***

Kadro ilgis N Polinomo eilė M	SNR= 30 dB	+ 65 dB triukšmas	+ 60 dB triukšmas	+ 55 dB triukšmas
N=250 M=9	0	7,2	24,3	56,7
N=250 M=10	0	6,3	23,4	47,7
N=250 M=10 Tiesinė skalė	0	11,2	29,7	63,9
N=500 M=9	0	9,9	21,6	52,2
N=500 M=10	0	5,4	22,5	45,0
N=500 M=10 Tiesinė skalė	0	8,1	27,9	51,3

LITERATŪRA

De WET, F.; WEBER, K.; BOVES, L.; CRANEN, B.; BENGIO, S. and BORLAND, H. (2004). Evaluation of formant-like features on an automatic vowel classification task. *The Journal of the Acoustical Society of America*, vol. 116(3), p. 1781–1792.

DELSARTE, P. and GENIN, Y.V. (1986). The Split Levinson Algorithm. *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-34 (3), p. 470–478.

FURUI S. (2001). *Digital Speech Processing, Synthesis, and Recognition*. Marcel Dekker, Inc.

GRUMADIENĖ, L; ŽILINSKIENĖ, V. (1997). Dažninis dabartinės rašomosios lietuvių kalbos žodynas. *Mokslo aidai.*

KABAL, P. and RAMACHANDRAN, R.P. (1986). The Computation of Line Spectral Frequencies Using Chebyshev Polynomials. *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-34 (6), p. 1419–1426.

LIPEIKA, A. L. (2005). Formantiniai požymiai atpažįstant kalbą. *Informacijos mokslai*, 34, p. 215–219.

LIPEIKA, A. L. (2007). Formantinių požymių išskyrimo metodai. *Informacijos mokslai*, 42–43, p. 201–206.

LIPEIKA, A. L. and LIPEIKIENĖ, J. (2008). On the Use of the Formant Features in the Dynamic Time Warping Based Recognition of Isolated Words. *Informatica*, 19 (3), p. 213–226.

TAMULEVIČIUS, G; LIPEIKA, A. L. (2003). Žodžių atpažinimo sistemos kūrimas. *Lietuvos matematikos rinkinys*, Spec. nr., t. 43, p. 292–296.

INVESTIGATION OF FORMANT FEATURES IN SPEECH RECOGNITION

Antanas Leonas Lipeika

S u m m a r y

The use of formant features in speech recognition is investigated in the paper. It was established that formant features can be used in speech recognition but recognition accuracy depends remarkably on the formant feature extraction method. The best recognition results were obtained when singular prediction polynomials were used for formant feature extraction. These polynomials can be calculated from parameters of linear prediction models of even or odd order. These polynomials can by symmetric or antisymmetric as well. Also it is important to investigate how results of speech recognition depends not only on choice of singular prediction polynomials but although on other parameters of the recognition system: frame length, number of used formants in recognition, frequency scale, used for representation of formant features. During the experiments it was defined that the best recognition results were obtained using 2 or 3 formants calculated from symmetric singular prediction polynomials. The experiments have shown that the most informative is the 2-nd formant. Contribution of the 1-st, 3-rd and 4-th formants is approximately similar, but higher formants are less resistant to white noise. Recognition results also depends on analysis frame length and frequency scale. The best results were obtained using 500 data points frame length and Mel frequency scale.