5 Spektrální sklon

Lenka Weingartová, Tomáš Bořil, Jitka Vaňková

Spektrální sklon (anglicky *spectral slope* nebo *spectral tilt*, řidčeji *spectral balance*) je termín, kterým se označuje spád křivky interpolující závislost amplitudy na frekvenci v řečovém spektru. Má na něj vliv mnoho různých řečových i neřečových faktorů, z nichž nejdůležitějším pro forenzní rozpoznávání je barva či kvalita hlasu. Stejně jako je např. F0 akustickým korelátem výšky hlasu, považuje se spektrální složení za akustický korelát jeho barvy. Protože barva hlasu je jednou z nejvýraznějších individuálních řečových charakteristik mluvčího, intenzivně se ve forenzní fonetice zkoumá.

S měřením spektrálního sklonu jsou nicméně spojeny určité metodologické potíže. Zatímco základní frekvenci můžeme změřit poměrně spolehlivě (viz 3. kapitolu) a její vztah k výšce hlasu je dnes už poměrně do hloubky prozkoumán, technické aspekty spektrální analýzy a zejména jejich následná interpretace pořád představují pro řečové odborníky výzvu. Ačkoliv jsme již o analýze spektra hovořili ve druhé kapitole v souvislosti s formanty, zde je pro důkladnější pochopení problematiky spektrálního sklonu nutný zevrubnější úvod.

5.1 Technické aspekty spektrální analýzy

Při analýze spektrálního sklonu vycházíme nejčastěji z digitálního záznamu řeči. Jedná se o signál, který je diskrétní v čase jako důsledek vzorkování a z důvodu kvantizace je také diskrétní v hodnotách. Vhodnou volbou vzorkovací frekvence při pořízení nahrávky ovlivňujeme maximální hodnotu frekvence, která bude v signále obsažena. Vzorkovací teorém udává, že vzorkovací frekvence musí být vyšší než dvojnásobek maximální frekvence, kterou chceme v signálu zaznamenat. Zatímco vzorkovací frekvence 8000 Hz (hertzy zde udávají počet vzorků za sekundu) je dostačující pro srozumitelný přenos obsahu promluvy telefonním kanálem, vyšší hodnoty jsou nutné pro uložení jemných detailů zabarvení lidského hlasu například při forenzních účelech identifikace mluvčího. Vždy se jedná o kompromis mezi velikostí nahrávky a množstvím informace, která zůstane v záznamu zachována. Frekvenční rozsahy zkoumaných pásem při výpočtu spektrálního sklonu popíšeme detailně v následujících podkapitolách.

Kvantizace digitálních signálů, neboli zaokrouhlování hodnot signálu na předem definované diskrétní hladiny, je nutným krokem k uložení řeči do paměti počítače. Jejím důsledkem jsou v podstatě náhodné odchylky skutečných a zaznamenaných hodnot vnímané jako aditivní šum, který ve skutečnosti v signálu přítomen nebyl. V dnešní době standardně používaná 16bitová hloubka lineární kvantizace je při dostatečně silném záznamu dostačující, kvantizační šum je i s kvalitními sluchátky prakticky neslyšitelný. Jiná situace nastává v případě, kdy má nahrávka nedostatečnou hlasitost a během zpracování je nutné ji zesilovat. Pokud to nahrávací technika umožní, je dobré volit během záznamu kvantizaci 24 bity pro maximální snížení úrovně kvantizačního šumu. Teprve po následném zesílení v programu pro editaci zvuku snížíme bitovou hloubku na 16 bitů, se kterými umí pracovat i běžné zvukové karty.

Zdaleka nejvhodnějším prostředím pro pořízení záznamu řeči je profesionálně vybavené studio. Důvodem je nejen samotná kvalita použité elektroniky, a tím minimalizace rušivých dopadů na signál, ale i konstrukce místnosti, která zajišťuje, že záznam neobsahuje ruchy pozadí ani odrazy řeči v samotném prostředí. Rezonanční frekvence například běžných kancelářských místností a netlumené odrazy od stěn a podlahy bez koberců mohou velice silně ovlivnit tvar výsledného spektra analyzované řeči.

Hlas je tvořen excitačním signálem, který je dále tvarován vokálním traktem a vyzařován do prostoru. Pokud chceme analýzami co nejlépe zachytit samotný hlas, musí být co nejméně ovlivněn impulsní odezvou (a tedy frekvenční charakteristikou) tohoto prostoru, ve kterém se následně šíří.

Mezi další nelingvistické faktory ovlivňující průběh záznamu patří vlastnosti mikrofonu a poloha mluvčího vůči němu. Mikrofon musí mít dostatečný frekvenční rozsah pro zachycení požadovaných frekvenčních složek. Důležitý je též jeho dynamický rozsah, jelikož může do nahrávky zanášet další šum při nízkých hlasitostech, či naopak může dojít k nelineárnímu zkreslení vlivem příliš silného akustického tlaku v důsledku malé vzdálenosti zdroje. Dále je třeba si uvědomit, že na intenzitu záznamu nemá vliv jen vzdálenost od úst mluvčího, ale také směr natočení mikrofonu. Tuto vlastnost udává tzv. směrová charakteristika mikrofonu. Zatímco různé zesílení porovnávaných nahrávek nehrají v dále definovaných způsobech výpočtu spektrálního sklonu roli (jak bude vysvětleno později), při zkoumání dlouhodobého spektra musíme mluvčího instruovat, aby v rámci této jedné vlastní nahrávky své pohyby vzhledem k mikrofonu minimalizoval, a to včetně natáčení hlavy.

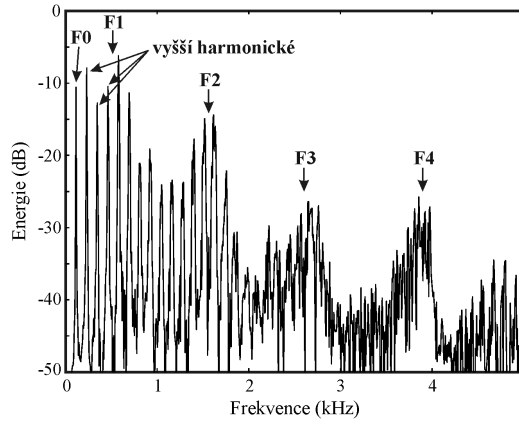
Pochopitelné je, že v mnoha případech si z čistě praktických důvodů nemůžeme dovolit komfort studiové nahrávky a laboratorních podmínek a musíme se spokojit se signálem různými způsoby degradovaným (okolní šum, filtrování GSM přenosem, atd.). I v takovém případě bychom však měli

mít na paměti, jaké prvky mohou ovlivňovat tvar spektra, a tím i spektrální sklon, a z nahrávek pak cíleně vybírat takové části, které jsou co nejméně zkreslené a obsahují minimum ruchů na pozadí. Doddington (1985: 1658) explicitně uvádí, že spektrální charakteristiky měřené pro forenzní účely jsou velmi citlivé na zkreslení signálu; French (1994: 175) doporučuje analyzovat pouze nahrávky srovnatelné kvality, a pokud tento případ nenastane, pak jednu z nich degradovat.

Nyní se zaměřme na jednotlivé aspekty spojené s měřením spektra ze získaného záznamu. Prostý digitální záznam zachycuje s určitou přesností časový vývoj akustického tlaku zaznamenaný v jednom místě prostoru. Jelikož mnoho tvarově odlišných průběhů takových záznamů může působit percepčně velice podobně, je výhodné časové průběhy převádět do spektrální oblasti, jež velice dobře souvisí s tím, jak zvuk skutečně vnímáme (kochlea ve vnitřním uchu provádí také spektrální rozklad).

Spektrum ve fourierovském smyslu nahlíží na signál jako na součet jednotlivých frekvenčních složek, konkrétně harmonických (sinusových) kmitů o různých frekvencích, amplitudách a fázích. Taková představa velice dobře odpovídá i tomu, jak vzniká řeč. Dle filtrové teorie produkce řeči jsou totiž jednou variantou zdroje excitačního signálu hlasivky, jejichž zvuk dále prochází vokálním traktem a na závěr je vyzařován do prostoru (Fant, 1960). Hlasivkové pulzy vytvářejí kvaziperiodický signál projevující se ve spektru jako tzv. základní harmonická složka s frekvencí, kterou označujeme Fo – jedná se o základní frekvenci kmitání hlasivek (viz obr. 5-1). Dále jsou přítomny složky označované jako vyšší harmonické, které by v případě ideálně periodického signálu byly celočíselnými násobky základní frekvence Fo. Příčinou těchto složek je kmitání hlasivek, které není jednoduchou sinusovkou, ale vlnou s mnohem komplikovanějším tvarem. Záleží na konkrétním mluvčím a konkrétní situaci, zda signál obsahuje více ostrých složek, tj. vyšších frekvencí, či naopak méně. Klesání amplitud jednotlivých harmonických s rostoucí frekvencí ve spektru se obvykle uvádí se sklonem zhruba 12 dB na dekádu (pokles amplitudy o 12 dB na desetinásobné frekvenci). Záleží však na témbru, hlasitosti a fonačním nastavení konkrétního mluvčího (viz dále oddíl 5.2).

Rozdílná nastavení vokálního traktu během řeči způsobují rezonance okolo různých frekvencí daných konkrétním nastavením, které pak tento signál (a jeho spektrum) dále tvarují. Vrcholky těchto rezonancí nazýváme formanty. Při produkci nazál pozorujeme i tzv. antiformanty, neboli propady energie ve spektru způsobené rezonancemi energie, která není vyzařovaná ven do prostoru.



Obrázek 5-1: Jednostranné spektrum vokálu [ɛ] znázorňující spektrální čáry základní frekvence Fo a vyšších harmonických, vrcholky vyšších hodnot amplitud určují formantové frekvence F1, F2, F3 a F4.

Takto vytvarovaný signál vychází ze rtů, nosu a stěn krku do prostoru, kde jej zachytí posluchač či záznamové zařízení. Díky tvaru těchto míst a také odlišné akustické impedanci vnitřního prostředí vokálního traktu (vlhkost, teplota) dochází k další modifikaci signálu, kterou ve spektru pozorujeme jako snížení sklonu zhruba o 6 dB na dekádu. Často se tedy v literatuře setkáváme s výslednou rámcovou hodnotou spektrálního sklonu 12 – 6 = 6 dB na dekádu, u které se ale uvádí, že je pouze přibližná a v čase neustále proměnlivá, závisející na mnoha okolnostech.

S určitou mírou zjednodušení tedy můžeme říci, že při změně melodie řeči se mění frekvenční poloha Fo a příslušných celočíselných násobků této hodnoty (vzdálenost jednotlivých čar ve spektru zobrazujících harmonické složky), zatímco poloha a tvar formantů (pomyslné kopce, které zesilují či zeslabují amplitudy jednotlivých harmonických) v případě shodné kvality vokálu zůstávají. Zajímavé je pak zkoumat celkový spektrální sklon složek, protože může vypovídat o mnoha skutečnostech, mimo jiné přispět právě i k rozpoznání identity mluvčího.

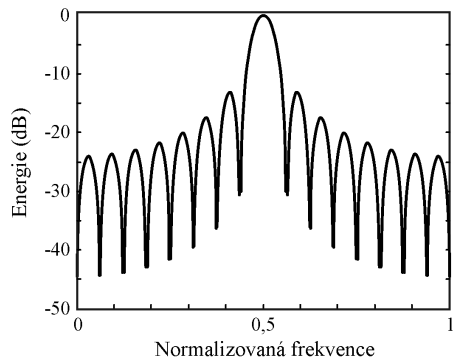
Doposud jsme v příkladu jako zdroj excitačního signálu uváděli kmitání hlasivek. Stejně tak se ale může jednat i o turbulentní proudění vzduchu při artikulaci neznělých frikativ či kombinaci takového náhodného proudění s kmitáním hlasivek u znělých frikativ. Všechny tyto jevy je možné ve spektru zobrazit. Při čistě turbulentním proudění vzduchu nebudou ve spektru harmonické složky, místo toho se šum projeví spektrem rozprostřeným přes velký rozsah frekvencí, kde ovšem můžeme spektrální sklon také měřit.

Jak jsme uvedli výše, fourierovská spektra chápeme jako znázornění situace, kdy je signál tvořen součtem nekonečně dlouho trvajících periodických sinu-

sovek o různých frekvencích, amplitudách a fázích. Problém nastává při praktickém použití diskrétní Fourierovy řady (anglická zkratka DFT), resp. výrazně rychlejšího algoritmu rychlé Fourierovy transformace (FFT), který poskytuje matematicky ekvivalentní výstup, ale počítaný výrazně efektivnějším způsobem. V reálné situaci totiž pracujeme pouze se segmentem signálu omezené délky, což má za následek nepříjemné zkreslení naměřených hodnot spektra.

V časové oblasti se dá vznik takového segmentu matematicky popsat jako vynásobení původního dlouhého signálu řeči krátkým segmentačním okénkem. V nejjednodušším případě tzv. pravoúhlého okna se jedná o samé jedničky v místě segmentu a nuly mimo segment, čímž fakticky dojde k „vyříznutí" segmentu z celého časového kontinua. Ukazuje se však, že tato operace má neblahý vliv na výsledné spektrum tohoto segmentu. Operaci násobení okénkem v časové oblasti odpovídá ve frekvenční oblasti konvoluce spektra originálního signálu se spektrem samotného okénka. Pakliže by původní signál byl pouhou jednoduchou sinusovkou, po oříznutí na velikost segmentu vychází ve spektru celá řada frekvenčních složek s různými amplitudami (viz obr. 5-2). Pozorujeme tzv. hlavní lalok v místě frekvence původní sinusovky s určitou šířkou, která je určena délkou segmentu. Kolem něj následuje celá řada vedlejších laloků. Čím bude délka analyzovaného segmentu větší, tím je šířka hlavního laloku užší, a tím pádem rozlišení frekvenční analýzy jemnější.

Pokud by originální signál obsahoval více harmonických složek, což je typický případ řeči, každá taková složka díky konvoluci na sebe nabalí hlavní a vedlejší laloky spektra segmentačního okénka a dochází tak k jevu, který se nazývá prosakování ve spektru (anglicky *spectral leakage*) a který může velice zkomplikovat následné analýzy spektra. Není totiž jasné, které spektrální vrcholky skutečně znamenají frekvenční komponent přítomný v původním signálu a které jsou pouze artefaktem matematické operace skryté za segmentací a výpočtem spektra.



Obrázek 5-2: Spektrum standardního pravoúhlého segmentačního okna. Uprostřed normalizované frekvenční osy se nachází hlavní lalok, který je z obou stran obalen vedlejšími laloky.

Abychom nepříjemný efekt spektrálního prosakování co nejvíce potlačili, při segmentaci cíleně násobíme okénky, která mají výhodnější spektrální vlastnosti. Vždy se jedná o kompromis mezi šířkou hlavního laloku (a tím možností rozlišit dvě blízké frekvenční složky v původním signálu) a mírou potlačení nežádoucích vedlejších laloků. Pro obecné aplikace zpracování signálů se doporučuje Hammingovo okno (Oppenheim & Schafer, 1989: 447), které má oproti pravoúhlému oknu jen o málo širší hlavní lalok, ale výrazně potlačuje vedlejší laloky. Pro potřeby zpracování řeči, kde harmonické složky bývají poměrně dosti vzdálené, se často využívá Gaussovo okno (Boersma & Weenink, 2014) se širším hlavním lalokem, ale ještě výraznějším potlačením vedlejších laloků (v grafickém znázornění šedivé škály při časovém vývoji spektra ve spektrogramu téměř nepozorovatelných).

Velice důležitou informací ovšem je, že pokud použijeme některé z těchto základních okének, dochází díky tvaru hlavního laloku ke zkreslení hodnoty amplitudy harmonických složek. Pokud potřebujeme ve spektru odečítat maxima spektrálních vrcholků a interpretovat je jako amplitudy složek v originálním signálu, což je případ některých ukazatelů zmiňovaných dále v této kapitole, je nutné použít k segmentaci tzv. „flat-top" okno (Gade & Herlufsen, 1987), které amplitudu složek zachovává, ovšem na úkor ještě širšího hlavního laloku než u Gaussova okna. V případě některých krátkých vokálů a hlubších mužských hlasů, kde jsou jednotlivé harmonické rozmístěny s malým rozestupem, ale můžeme narazit na problém šířky hlavního laloku, který pak způsobí rozmazání jednotlivých složek přes sebe; v takovém případě je spektrální rozlišení nevyhovující.

V následujícím textu se několikrát zmíníme o energii vybraného pásma spektra. Spektrální sklon pak může být definován jako poměr energií dvou pásem spektra. Spektrum vypočtené pomocí metody FFT obsahuje komplexní čísla pro jednotlivé frekvence. Každé komplexní číslo v sobě obsahuje amplitudu a fázi konkrétní harmonické složky o dané frekvenci. Energetické spektrum počítáme pouze z amplitudové části spektra, které vypočteme jako absolutní hodnoty komplexních čísel. Vzhledem k tomu, že záznam signálu řeči nebývá kalibrován ke vztažné hodnotě akustického tlaku, uvádíme tyto hodnoty bez jednotek. Umocněním amplitudy na druhou obdržíme hodnotu energie pro danou harmonickou složku. Výsledkem FFT je tzv. oboustranné spektrum, které z důvodu možnosti transformace i komplexních signálů, jež ale v našem oboru nevyužijeme, zavádí i fiktivní záporné frekvence. Praktickým závěrem z této skutečnosti je pro nás to, že pokud zobrazíme takové amplitudové či energetické spektrum z našich reálných signálů, obdržíme vždy zrcadlově symetrické spektrum podle počátku, tedy levou (záporné frekvence) a pravou polovinu (kladné frekvence). Pokud bychom chtěli pracovat se skutečně celkovou energií signálu, použili bychom jednu polovinu energetického spektra a všechny hodnoty vynásobili dvěma. Vzhledem

k tomu, že nám jde ale jen o poměr energií mezi pásmy, násobit dvěma není nutné, protože by se číslo dvě v poměru vykrátilo.

Z energetického spektra je pak možné vypočítat jednoduše souhrn energie ve vybraných pásmech. Standardně se energie digitálních signálů udává v dBFS (hodnoty vztažené k maximálnímu rozsahu kvantování), označované nejčastěji pouze zkratkou dB, které se vypočtou dle vztahu

$$energie_{dB} = 10 \times \log_{10}(energie).$$

V následujícím textu budeme uvažovat vždy již energii vyjádřenou v dB, a proto budeme některé výpočty spektrálního sklonu, které značí poměr energií dvou pásem, uvádět jako rozdíl dvou hodnot energií vyjádřených v dB. Praktické důsledky hodnot v dB jsou zřejmé – celkové zesílení či zeslabení intenzity nahrávky (matematicky násobení celého signálu konstantou) se projeví v dB pouze vertikálním posunem celého spektra nahoru či dolů. Poměr energií, což je rozdíl energií vyjádřených v dB, způsobí vykrácení resp. odečtení zesilující konstanty. Opticky i početně zůstává spektrální sklon stejný.

5.2 Lingvistické koreláty spektrálního sklonu

Ukazuje se, že na spektrální sklon má kromě osoby mluvčího a nelingvistických faktorů zmíněných výše vliv i několik faktorů lingvistických, přičemž situace není zatím prozkoumána v celé úplnosti. Spektrální sklon je všeobecně považován za akustický korelát *kvality* nebo *barvy hlasu*, *tembru* (někdy se můžeme setkat i s označením *rejstřík*, které ovšem není jednoznačně definováno a může zahrnovat jen některé druhy hlasových charakteristik). Těmito termíny se obvykle označuje typické zabarvení hlasu mluvčího, percepční abstrakce z krátkodobých artikulačních charakteristik, které mluvčí používá k přenosu lingvistických i nelingvistických informací (Laver, 1980: 1). Barva hlasu se dlouho popisovala čistě percepčně, podle dojmu, který hlas v posluchači vyvolává (např. „světlý“ vs. „tmavý“ hlas). Jak bylo nicméně ukázáno výše, současné řečové technologie umožňují zkoumat spektrální složení hlasu a zjišťovat, které akustické charakteristiky jsou příčinou těchto dojmů.

Plošší spektrum, tedy větší množství energie ve vyšších frekvencích (nižší spektrální sklon), vykazuje hlas, který bychom mohli popsat jako řezavý, napjatý či drsný (Laver, 1980). Stejný efekt má na spektrum také falzet nebo třepená fonace (Monsen & Engebretson, 1977: 988; Hammarberg, Fritzell, Gauffin, Sundberg & Wedin, 1980: 446), ke které dochází při zpomalení kmitání hlasivek tak, že jsou rozeznatelné jednotlivé pulzy a není již udržena pravidelnost kmitání. Spektrální sklon se také snižuje se zvyšováním mluvního úsilí, které přímo souvisí se zvyšováním hlasitosti a projevuje se opět nárůstem energie ve vyšších frekvencích (Doddington, 1985: 1659; Sluijter & van Heuven, 1996: 2482n.). Zvyšování mluvního úsilí je pravděpodobně

také příčinou nižšího spektrálního sklonu naměřeného při tzv. Lombardově jevu (viz např. Bořil & Hansen, 2010), ke kterému dochází při mluvení v nepříznivých podmínkách, typicky v hlučném prostředí, kdy mluvčí zvyšuje základní frekvenci, mluvní úsilí a zpřesňuje artikulaci.

Zkoumají se též spektrální vlastnosti hlasu v souvislosti s emocemi – snížení spektrálního sklonu je obvykle spojováno se vztekem či strachem (Banse & Scherer, 1996; Tamarit, Goudbeek & Scherer, 2008). Naopak hlas klidný, tichý, dyšný či jemný vykazuje spektrální sklon největší, čili ve vyšších frekvencích ubývá energie výrazně rychleji.

Spektrální sklon je také uváděn jako korelát lingvistické prominence, přízvučnosti slabiky. Ukazuje se, že v jazycích, jako je angličtina nebo holandština, má přízvučná slabika sklon menší oproti nepřízvučné (Sluijter & van Heuven, 1996; Sluijter, van Heuven & Pacilly, 1997; Campbell & Beckman, 1997 nebo Heldner, 2001, ale srv. Monsen & Engebretson, 1977: 991). Prominence je nicméně komplexní výsledek souhry několika faktorů (např. zvýšení základní frekvence, zvýšení mluvního úsilí nebo prodloužení trvání hlásky) a role spektrálního sklonu a jeho koordinace s ostatními faktory není dosud uspokojivě popsána.

Protože spektrální sklon do velké míry závisí na vlastnostech vokálního traktu mluvčího a na jeho charakteristickém způsobu fonace, je využíván také pro rozpoznávání mluvčího, viz např. Nolan (1983: 130nn.) nebo Doddington (1985). Jednotlivým metodám měření spektrálního sklonu a jejich schopnostem rozlišit od sebe mluvčí budou věnovány následující podkapitoly.

5.3 Dlouhodobé ukazatele spektrálního sklonu

Dlouhodobé ukazatele spektrálního sklonu se odvíjejí od tzv. dlouhodobého průměrného spektra, LTAS (z anglického *Long-Term Average Spectrum*). LTAS je metoda akustické analýzy, která poskytuje informace o spektrálním rozložení energie v řečovém signálu za delší časový úsek (Löfqvist, 1986). Jelikož je řečový signál produktem zdroje i filtru vokálního traktu, který se liší pro jednotlivé hlásky, zahrnutím většího vzorku docílíme toho, že krátkodobé změny dané právě vlivem segmentálním se zprůměrují. Výsledné spektrum tedy reflektuje celkové přispění jak zdroje, tak vokálního traktu ke kvalitě hlasu mluvčího a není ovlivněno segmentálními rozdíly v řečovém materiálu (Nordenberg & Sundberg, 2003; Master, De Biase, Pedrosa & Chiari, 2006).

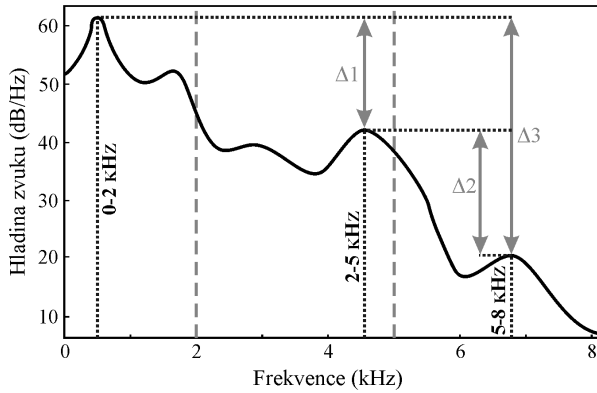
Dosavadní výzkum ukázal, že LTAS je citlivé na různé kvality hlasu a používá se proto jako objektivní doplnění kvality vnímané (Hammarberg et al., 1980; Laver, 1980; Kitzing, 1986). LTAS se zatím ukázalo jako užitečný ukazatel rozdílů mezi mužskými a ženskými hlasy (Mendoza, Valencia, Munoz & Trujillo, 1996; White, 2001; Nordenberg & Sundberg, 2003), rozdílů ve věku

(Linville, 2002; da Silva, Master, Andreoni, Pontes & Ramos, 2011) nebo mezi trénovanými a netrénovanými hlasy – např. Sundberg (1987) identifikoval vrcholek mezi 2,8 a 3,4 kHz, když analyzoval hlasy zpěváků pomocí LTAS. Tento vrcholek, který byl pojmenován „pěvecký formant" (*singer's formant*), vzniká seskupením F3, F4 a F5 a je spojen s vnímáním zvučných hlasů. Tento poznatek byl později potvrzen také Leinem (1993), který studoval hlasy herců, a mluví o řečnickém či hereckém formantu (*„speaker's/actor's formant"*). V neposlední řadě se LTAS ukázalo jako vhodný nástroj při diagnostice dysfonických hlasů (Hammarberg et al., 1980) nebo při hlasové nápravě po terapii (Kitzing & Åkerlund, 1993; Tanner, Roy, Ash & Buder, 2005).

Co se týče forenzního využití, vektor dlouhodobého spektra tvoří část Hollienova poloautomatického systému rozpoznávání mluvčího SAUSI (*Semi-Automatic Speaker Identification System*; Hollien, 2002). Tento vektor je považován za citlivý na identitu mluvčího, a to i za nepříznivých podmínek jako např. přítomnost šumu, omezené pásmo nebo stres (Hollien, 2002: 162).

Bylo navrženo několik parametrů, které LTAS kvantifikují. Tyto parametry vypovídají o celkovém sklonu spektrální obálky, a tedy nějakým způsobem také o kvalitě hlasu (Hammarberg et al., 1980; Leino, 1993, citováno z Master et al., 2006). Spektrální sklon je obvykle výsledkem porovnání určitých spektrálních vrcholků, nebo se vyjadřuje jako rozdíl množství energie v určitých frekvenčních pásmech.

Jak již bylo zmíněno výše, z psychoakustického hlediska by měly zvučné hlasy oproti hlasům tlumeným či drsným vykazovat v LTAS více energie ve vyšších harmonických, tj. méně strmý spektrální sklon (Löfqvist, 1986), což mnohé studie opravdu ukazují. Jedním z prvních obecně přijatých a dále rozpracovávaných způsobů měření dlouhodobého spektrálního sklonu se stala metoda Hammarbergové a jejích kolegů (1980), kteří vyjadřovali spektrální sklon jako rozdíl energií nejsilnějších vrcholků ve třech frekvenčních pásmech LTAS (0–2, 2–5 a 5–8 kHz, viz obr. 5-3) s automatickou eliminací neznělých hlásek. Ve svých experimentech našli například významnou korelaci mezi hlasy vnímanými jako dyšné a spektrálním sklonem, a to konkrétně ve strmějším úpadku energie mezi pásmy 0–2 kHz a 2–5 kHz. Tento index byl poté nazván index Hammarbergové (*Hammarberg index*). Podle experimentálních výsledků reflektuje nejen rozdíl mezi různými kvalitami hlasu (Hammarberg et al., 1980), ale také různými řečovými styly (Monzo, Alías, Iriondo, Gonzalvo & Planet, 2007). Nižší koncentrace energie v oblasti nad pásmem prvního formantu a vyšší koncentrace energie nad 5 kHz byla spojena s dyšnou fonací či hypofunkčním hlasem rovněž v dalších studiích (Soyama, 2005, citováno z Master et al., 2006).



Obrázek 5-3: Měření spektrálního sklonu jako rozdílu maximálních energií LTAS v pásmech 0–2 kHz, 2–5 kHz a 5–8 kHz.

Výzkumníci se od té doby snaží najít další vztahy mezi vrcholky či frekvenčními oblastmi spektrálního sklonu, které by ho kvantifikovaly, a reflektovaly tak vnímanou kvalitu hlasu. Jedním z nejčastěji používaných je *index alfa* (α) (Frøkjaer-Jensen & Prytz, 1976), který také vychází z LTAS a stejně jako index Hammarbergové pramení ze snahy o nalezení spektrálních charakteristik hlasových poruch. Počítá se jako rozdíl energií (tedy již ne maxim) nad a pod hranicí 1000 Hz, konkrétně (1–5 kHz) minus (0–1 kHz). Index α se ukázal jako užitečný pro rozlišení kvality hlasu v několika studiích (např. Löfqvist, 1986; Sundberg & Nordenberg, 2006; Leino, 2009). Výsledky dalších studií týkajících se indexu α a patologických hlasových kvalit sice nejsou příliš konzistentní, nicméně Kitzing (1986) zařazuje tento ukazatel mezi několik schopných odlišit od sebe zdravé hlasy při různých typech fonace, ačkoliv nalezené rozdíly dosahovaly jen řádu několika procent (Kitzing, 1986: 481). Sundberg a Nordenbergová (2006) potvrzují vztah mezi indexem α a hlasitostí (přesněji mluvním úsilím – jejich mluvčí hovořili do zvyšujícího se hluku); čím hlasitěji se subjekt snaží mluvit, tím nižší má spektrální sklon. Jejich materiál zahrnoval i neznělé hlásky.

Index α navíc sloužil jako základ pro parametr další, a to *Kitzingův index* (Kitzing, 1986). Kitzing ve své studii ukázal, že převrácená α, tedy rozdíl energií pod a nad 1000 Hz, je také spolehlivým ukazatelem kvality hlasu. Omezuje se však jen na frekvence do 2 kHz a je tedy roven rozdílu energie v pásmu 0–1 kHz a 1–2 kHz.

Kromě indexů Hammarbergové, α a Kitzingova indexu byly navrženy i další kvantifikace spektrálního sklonu, jako např. rozdíl energií v pásmu 0–1 kHz a 1–6,5 kHz či 0–1 kHz a 1–20 kHz, která zahrnuje všechny frekvence slyšitelné lidským uchem (Sergeant & Welch, 2008). Také Ternström (2008)

zkoumal pomocí LTAS rozložení energie v oblastech nad 5 kHz. Přestože opomenutí energie nad 5 kHz nepostihne srozumitelnost řeči, jelikož většina energie je soustředěna právě v oblasti do 5 kHz, tento rozsah je slyšitelný a důležitý pro percepci. Ternströmova data ukazují, že kontura LTAS je pro mluvčí do jisté míry specifická i ve frekvenční oblasti od 5 do 20 kHz.

Jiným způsobem porovnávali spektrální sklon na velkém korpusu Kochanski, Grabeová, Coleman a Rosner (2005) – sklon spektrální obálky je v jejich experimentu kvantifikován regresní křivkou, která modeluje spektrální obálku mezi 500 a 3000 Hz, ovšem normalizovanou vzhledem k percepční odezvě, v intervalech po 1 Barku. Tamarit et al. (2008) potvrzují užitečnost regresní křivky (používají lineární a exponenciální regresi) pro rozpoznání změny spektrálního sklonu při emočně expresivní řeči.

Přestože užitečnost dlouhodobých ukazatelů jako právě LTAS spočívá ve vyrušení vlivu jednotlivých hlásek, čímž získáme představu o průměrných hodnotách pro daného mluvčího, nejsou ani dlouhodobé ukazatele netečné vůči nejrůznějším faktorům (Rose, 2002: 59). Např. Nordenbergová a Sundberg (2003) ve své studii ukázali, že srovnatelnost dat vyslovených různou hlasitostí je diskutabilní z toho důvodu, že nárůst amplitud v jednotlivých pásmech spektra není vzhledem k celkové změně intenzity lineární – ve středních frekvencích (1500–3000 Hz) je výraznější než v nižších. To motivovalo autory k detailnějšímu zkoumání vlivu produkční hlasitosti na LTAS. V pozdější studii (Sundberg & Nordenberg, 2006) ukázali, že nárůst sice lze aproximovat určitými funkcemi, ale existuje zde variabilita mezi jednotlivci.

Otázka spolehlivosti LTAS byla prozkoumána také ve studii Löfqvistově (1986), která naznačuje, že variabilita LTAS, které zde bylo kvantifikováno pomocí indexu α (tedy rozdílem energie v pásmu 0–1 kHz a 1–5 kHz), může být značná dokonce v rámci jednoho mluvčího. V jeho studii měli mluvčí ke konci dne značně strmější spektrální sklon než na začátku dne.

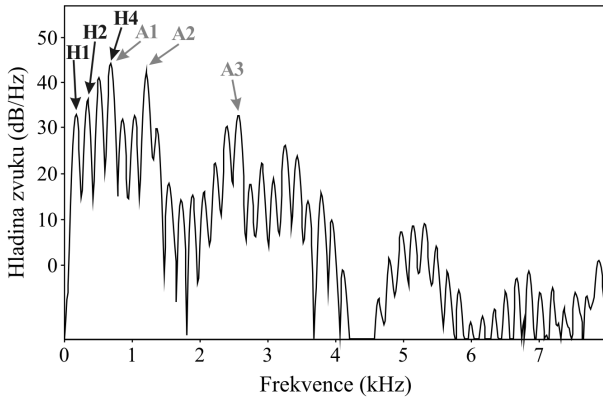
Při použití pro forenzní účely je tedy nutné mít všechny tyto faktory, které mohou mít na LTAS vliv, na paměti.

5.4 Krátkodobé ukazatele spektrálního sklonu

Měříme-li spektrální sklon krátkodobý, znamená to, že vykreslujeme průměrné spektrum jen krátké části zvukového signálu, typicky jedné hlásky, případně její části. Zde hraje velkou roli identita zkoumaného segmentu – zatímco frikativy vykazují spektrální sklon blížící se nule, u vokálů pozorujeme charakteristický úbytek energie směrem k vyšším frekvencím. Dlouhodobá spektra tyto rozdíly průměrují, čímž však dochází ke ztrátě velkého množství informací, které mohou obsahovat idiosynkratické rysy a také pomoci rozlišit mluvčí. Je sice pravda, že měření krátkodobých ukazatelů bylo využíváno především ke zkoumání slovního přízvuku, mnozí autoři však

zmiňují jejich variabilitu mezi mluvčími, proto je praktické prozkoumat i jejich forenzní využití.

Pro forenzní účely se zdají být slibné parametry, které reflektují glotální nastavení. Jde o skupinu parametrů, které porovnávají amplitudy různých akustických událostí (viz obr. 5-4). Typicky je vztažena amplituda první harmonické (H1) k nějaké další. Nejčastěji se v literatuře objevuje parametr H1-H2 (tedy rozdíl amplitud prvních dvou harmonických vyjádřených v dB), H1-A1 (rozdíl amplitudy první harmonické a nejsilnější harmonické v oblasti prvního formantu) a H1-A3, kde A3 je nejsilnější harmonická v oblasti třetího formantu (Hanson, 1997). Objevují se však i další metriky, jako např. H2-H4 (druhá harmonická minus čtvrtá harmonická) a H1-A2 (první harmonická minus nejsilnější harmonická druhého formantu) (Garellek, Samlan, Kreiman & Gerratt, 2013).



Obrázek 5-4: Vokalické spektrum s vyznačenými amplitudami harmonických, které se používají pro výpočet glotálních parametrů. Převzato z: Vaňková et al. (2014).

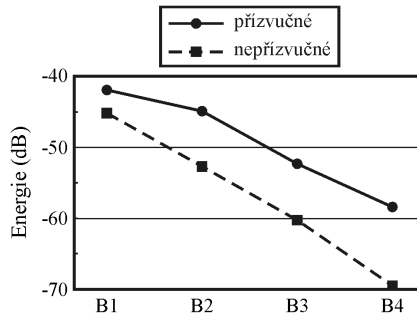
Jak již bylo řečeno, tyto parametry vypovídají o glotálních charakteristikách – např. H1-H2 koreluje s poměrem otevření (Hanson, 1997), což je poměr trvání fáze otevření a zavření hlasivek v rámci jedné periody. H1-A1 pak vypovídá o šířce pásma prvního formantu, a tedy míry, do jaké glotis zůstává během fáze zavírání otevřena, a H1-A3 vypovídá o strmosti glotálního pulzu a délce fáze zavírání (Hanson, 1997).

Aby bylo možné hodnoty parametrů pro různé vokály porovnávat napříč mluvčími, je nutné provést normalizaci vlivu vokálního traktu (Iseli, Shue & Alwan, 2007). Takto opravené parametry se pak značí hvězdičkou, tedy H1*-H2*, H1*-A1*, H1*-A2*, H1*-A3* and H2*-H4*.

Nedávná studie zaměřená na analýzu variability těchto parametrů ukázala, že především H1*-H2*, H1*-A1* a H1*-A2* vykazují nízkou variabilitu v rámci jednoho mluvčího a vysokou mezi mluvčími, což je pro forenzní

účely záhodné. Dodejme ještě, že když měla lineární diskriminační analýza na základě těchto pěti parametrů rozlišit 6 mluvčích, bylo dosaženo lepších výsledků, než když byli mluvčí rozlišováni na základě statických hodnot frekvencí formantů (F1–F4), které se pro tyto účely jinak běžně používají (Vaňková & Skarnitzl, 2014). Navíc se zdá, že tyto parametry jsou celkem robustní také v méně příznivých podmínkách, jako např. při komprimaci nahrávek při mobilním přenosu (Vaňková, Bořil & Skarnitzl, 2014; viz také 7. kapitolu).

Jiný typ krátkodobého spektrálního sklonu měřili Sluijterová a van Heuven (1996), kteří se snažili potvrdit hypotézu, že spektrální sklon je akustickým korelátem přízvučnosti vokálu. Spočítali energie ve čtyřech frekvenčních pásmech (0–0,5 kHz, 0,5–1 kHz, 1–2 kHz a 2–4 kHz), která byla určena tak, aby zahrnovala postupně oblasti základní frekvence a prvních tří formantů. Spektrum měřili na jediném vokálu, vždy v bodě nejvyšší hodnoty prvního formantu. Konkrétní příklad měření uvádí obrázek 5-5: hodnoty pro nepřízvučný vokál [ɛ] jsou vyznačeny čárkovaně, pro přízvučný pak plnou čarou. Vidíme, že nepřízvučný vokál má větší spektrální sklon, tedy méně energie ve vyšších frekvenčních pásmech.



Obrázek 5-5: Měření spektrálního sklonu v přízvučném (plná čára) a nepřízvučném (čárkovaně) vokálu [ɛ] jako energií ve čtyřech pásmech: 0–0,5 kHz, 0,5–1 kHz, 1–2 kHz a 2–4 kHz. Hodnoty pro jednoho mluvčího mužského pohlaví pocházejí z korpusu využitého ve studii Weingartové a Volína (2014).

Pro diskriminaci přízvučných a nepřízvučných vokálů bylo rovněž využito několik metod založených na indexu α, které porovnávají rozdíly energií ve dvou frekvenčních pásmech s různými hranicemi (Prieto & Ortega-Llebaria, 2006). Tuto metodu převzali také Volín a Zimmermann (2011), ovšem se dvěma zásadními změnami. Za prvé, z výpočtu vyřadili frekvenční pásmo F0 na základě úvahy, že masivní energie základní frekvence by přebila jemnější spektrální detaily. Vyřazeno bylo též pásmo F2 – autoři předpokládali, že F2 kóduje primárně identitu hlásky, a pokud by toto lokální maximum spadalo pokaždé do jiného pásma, mohlo by to výrazně ovlivnit měření spek-

trálního sklonu. Mezi dvěma měřenými pásmy tedy vznikla mezera.[1] Tato metoda byla při všech nastaveních schopna úspěšně rozpoznat přízvučné a nepřízvučné vokály tří mluvčích mužského pohlaví, přičemž nejúspěšnější byla pásma 350–1100 a 2300–5500 Hz. Stejná metoda s výše zmíněnými nejúspěšnějšími frekvenčními pásmy byla následně využita ve studii Volína, Weingartové a Skarnitzla (2013) k rozpoznání šva tří českých a tří rodilých mluvčích angličtiny ženského pohlaví.

Jak si autoři první studie povšimli, zásadní otázkou při měření krátkodobého (stejně jako dlouhodobého) sklonu zůstává volba hranic frekvenčních pásem. Ze stávajících výsledků výzkumu lze předpokládat, že ne všechna frekvenční pásma jsou pro kvantifikaci spektrálního sklonu relevantní – záleží zejména na výzkumné otázce, kterou si klademe.

Zda je relevantní pásmo F0, je otázka, která si zaslouží další zkoumání. Výsledky Sluijterové a van Heuvena (1996) naznačují, že ano, pro kvantifikaci přízvuku je potřeba zahrnout i pásmo základní frekvence, jinak se ostatní pásma liší pouze relativní intenzitou, nikoliv sklonem křivky (viz obr. 5-5). Pro zkoumání patologických hlasových kvalit se podle výsledků Hammarbergové et al. (1980) zdá být toto pásmo také velmi důležité. Lepších výsledků při zahrnutí pásma F0 dosáhli i Weingartová a Volín (2014) ve studii, která bude zmíněna níže.

Otázka vysokých frekvencí při měření krátkodobého spektrálního sklonu opět není dosud uspokojivě zodpovězena – v závislosti na kvalitě nahrávky můžeme mít k dispozici spektrum frekvencí například do 16 kHz, kde se ale již nenacházejí řečové informace.

V relevantní literatuře se horní hranice pásem dosti liší, a jen málokteří autoři svou volbu vysvětlují. Hammarbergová (Hammarberg et al., 1980) měřila pásma do 8000 Hz, což bylo pravděpodobně maximum dané tehdejšími technologickými možnostmi. Maximum 8000 Hz využili také Eriksson, Thunberg & Traunmüller (2001), kteří jej porovnávají s pásmem F0, svůj ukazatel nazývají *spektrální emfáze* a kvantifikují jím prominenci ve švédštině. Stejně vysokou hranici zvolili ještě Tamarit et al. (2008) s vysvětlením, že i v těchto vysokých frekvencích mohou ležet zajímavé informace. Dále byly využity hranice 6000 Hz (Sundberg a Nordenbergová, 2006), 5000 Hz (Frøkjaer-Jensen & Prytz, 1976; Kitzing, 1986; Banse & Scherer, 1996), 4000 Hz zvolili Sluijterová a van Heuven (1996) a Boersma a Kovačićová (2006), 2000 Hz použil Kitzing (1986).

První forenzně zaměřená studie týkající se krátkodobého spektrálního sklonu v českém prostředí ověřila úspěšnost výše zmíněných měření spek-

[1] Měřit nesousedící frekvenční pásma vyzkoušel již Kitzing (1986: 480), 300–800 Hz a 1500–2000 (3000) Hz, ovšem nevěnuje jim ve svém článku žádnou pozornost, ani nevysvětluje motivaci hranic vybraných pásem. Ve srovnání s indexem α dopadly o něco hůře.

trálního sklonu jako rozdílu energií v pásmech 350–1100 a 2300–5500 Hz při rozpoznání čtyř mluvčích mužského pohlaví (Weingartová & Volín, 2013). Zároveň testovala různé možnosti konkrétního výpočtu spektrálního sklonu z energií dvou frekvenčních pásem, a kromě toho navíc ukazatel sešikmení spektra, který se ovšem ukázal pro diskriminaci mluvčích spíše nevhodný.

5.4.1 Experiment[2]

Výsledky první studie naznačují, že i krátkodobá spektra jsou citlivá na identitu mluvčího – platí to ale pro všechny ukazatele? Tuto otázku se pokusil zodpovědět novější výzkum (Weingartová & Volín, 2014), kde bylo prověřováno, jak jsou různé ukazatele spektrálního sklonu ovlivněny identitou mluvčího, identitou hlásky a slovním přízvukem. Autoři využili laboratorní, vysoce kontrolovanou řeč, aby bylo možné všechny faktory podchytit. Pro naše účely postačí uvést, že mluvčí nijak neměnili hlas, ani se na něj nesoustředili, protože hlavním úkolem bylo vyslovit na správném místě přízvučnou slabiku. Analyzovaný korpus se sestával z nahrávek dvanácti mluvčích mužského pohlaví ve věku od 20 do 30 let, analyzovaných vokálů bylo celkem 1536.

Ověřovány byly následující ukazatele:

- Index Hammarbergové (HI) vypočítaný jako rozdíl mezi amplitudovými maximy ve frekvenčních pásmech 0–2 kHz a 2–5 kHz.
- Index α, který je výsledkem rozdílu energií v pásmech 0–1 kHz a 1–5 kHz (horní hranice 5000 Hz byla vzata v úvahu jako nejčastěji používaná v literatuře).
- Lineární regrese (LinReg) je přímka, která aproximuje sklon spektra ve frekvenčním pásmu 500–3000 Hz (podle: Kochanski et al., 2005). Byla využita logaritmická frekvenční škála a robustní aproximace.
- Lineární aproximace metody Sluijterové a van Heuvena (1996) – 4B LinFit. Lineární aproximace je potřeba z toho důvodu, abychom čtyři hodnoty ze čtyř pásem (viz obr. 5-5 výše) převedli na jedinou, která popisuje sklon. Jde o pásma 0–0,5 kHz, 0,5–1 kHz 1–2 kHz a 2–4 kHz.
- Rozdíl energií v pásmech 350–1100 a 2300–5500 Hz (BgNoF0), kde není zahrnuto pásmo F0 ani F2 (podle Volína a Zimmermanna, 2011).
- Rozdíl energií v pásmech 0–1100 a 2300–5500 Hz (Bg), tedy se zahrnutou F0.
- Rozdíl energií s pohyblivým pivotem na hodnotě druhého formantu (podle: Tamarit et al., 2008), bez F0 v pásmech 350–F2 a F2–5500 Hz (BpNoF0).
- Tentýž rozdíl, ale se zahrnutou F0 v pásmech 0–F2 a F2–5500 Hz (Bp).
- Spektrální emfáze (SpEm) vypočítaná jako rozdíl energií v celém spektru (od 0 Hz do Nyquistovy frekvence, což je v našem případě 16 kHz) a v pásmu 0–1.43×F0 (podle Erikssona et al., 2001).

[2] Založeno na studii Weingartové a Volína (2014).

Metoda	Faktor					
	Mluvčí		Vokál		Přízvuk	
	F	*p*	*F*	*p*	*F*	*p*
HI	24,3	<0,001	361,3	<0,001	19,9	<0,001
α	11,7	<0,001	712,7	<0,001	14,5	<0,001
LinReg	5,7	<0,001	1167,7	<0,001	3,5	nevýzn.
4B-LinFit	23,5	<0,001	470,2	<0,001	24,8	<0,001
BgNoF0	20,4	<0,001	400,5	<0,001	4	0,04
Bg	24,9	<0,001	414,3	<0,001	22,5	<0,001
BpNoF0	15	<0,001	194,3	<0,001	4,4	0,03
Bp	19,4	<0,001	339,7	<0,001	27	<0,001
SpEm	115,6	<0,001	19,4	<0,001	6	0,01
A1*-A2*	7,1	<0,001	257,3	<0,001	4,8	0,03

Tabulka 5-1: Výsledky tří jednofaktorových analýz rozptylu s faktory Mluvčí, Vokál a Přízvuk, převzato a upraveno ze studie Weingartové a Volína (2014: 7). Popis jednotlivých metod viz text.

Z tabulky 5-1 je vidět, že všechny zkoumané metody jsou citlivé na mluvčího – faktor byl vždy vysoce významný na úrovni $p < 0{,}001$. To znamená, že jednotliví mluvčí se od sebe liší v hodnotách spektrálního sklonu, nezávisle na vokálu či jeho přízvučnosti. Nejlepších výsledků dosahuje ukazatel spektrální emfáze (SpEm), nejhorších pak lineární regrese (LinReg), ale i u té je výsledek stále vysoce významný. Spektrální emfáze od sebe z možných 66 dvojic mluvčích rozlišila 54. Druhý nejúspěšnější ukazatel, rozdíl dvou pásem s mezerou v pásmu F2 (Bg), od sebe rozlišil 35 dvojic mluvčích.

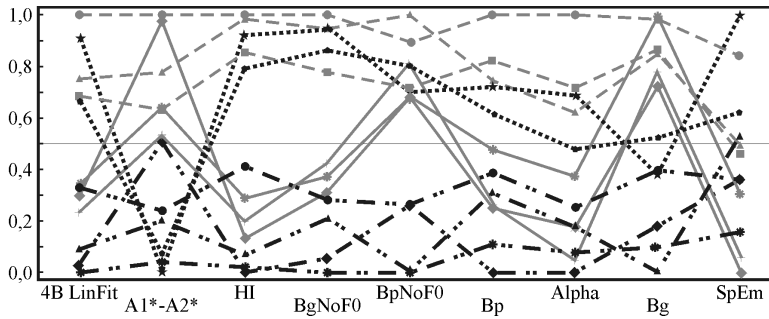
Ovšem neměli bychom opominout výsledky i u ostatních faktorů. Vysoká čísla v levém sloupci u faktoru Vokál znamenají, že ukazatel je také citlivý na identitu vokálu, a to v některých případech až řádově více než na mluvčího. Zde opět vychází nejlépe ukazatel spektrální emfáze, který je nejméně citlivý na vokály. Naopak lineární regrese spektra nejlépe postihuje identitu vokálu a vůbec nepostihuje jeho přízvučnost.

Ve světle těchto výsledků by se zdálo, že je nutné kontrolovat identitu vokálu, pokud porovnáváme různé mluvčí mezi sebou. Bohužel to v sobě skrývá nevýhodu – s vyřazením některých vokálů se výrazně snižuje množství případů a tím i úspěšnost ukazatelů. Zde jsme měli k dispozici průměrně 128 vokálů na mluvčího. Vybereme-li pouze jeden vokál, klesne počet případů na přibližně 25 na mluvčího a statistické testy tak ztratí svou sílu. Každopádně

bude-li zastoupení vokálů ve vzorku porovnávaných mluvčích nerovné (tj. například jeden mluvčí pronese víc [ɪ] a druhý víc [ɛ]), lze očekávat, že nalezené rozdíly budou reflektovat spíše tento rozdíl než rozdíl mezi hlasy mluvčích.

Zároveň se ovšem nemusíme omezovat pouze na jeden ukazatel spektrálního sklonu. Různé metody výpočtu výše zmíněných ukazatelů mají za výsledek to, že shlukují mluvčí různě – a tedy například dvojice mluvčích nerozlišitelná jedním z ukazatelů může být odlišitelná jiným.

Tuto situaci zobrazuje graf na obrázku 5-6, ve kterém jsou hodnoty pro všech dvanáct mluvčích převedeny do intervalu <0,1>, tak, že mluvčí s nejvyšší hodnotou má přiřazeno 1 a s nejnižší 0. Ostatní mluvčí jsou pak škálováni relativně k nim. Metoda lineární regrese byla z této analýzy a priori vyřazena jako nejméně užitečná. Pro co největší omezení variability dané identitou hlásky jsou v grafu zobrazeny pouze hodnoty pro vokál [ɛ].



Obrázek 5-6: Normalizované hodnoty devíti ukazatelů spektrálního sklonu pro 12 mluvčích a vokál [ɛ]. 1 je přiřazena mluvčímu s nejvyšší hodnotou, 0 s nejnižší. Mluvčí jsou rozděleni do čtyř skupin (naznačeny různými druhy čar). Převzato z: Weingartová & Volín (2014).

Z obrázku 5-6 jsou patrné dvě důležité skutečnosti – kromě toho, že ukazatele skutečně shlukují různé mluvčí různě, lze navíc rozeznat čtyři skupiny mluvčích, které se chovají více či méně podobně napříč různými metodami (jsou znázorněny různými druhy čar). Lze hypotetizovat, že mluvčí spadající do stejné skupiny mají podobné spektrální vlastnosti hlasu, a tedy by mohli i podobně „znít". To by ovšem bylo nutné ověřit percepčním testem, který nebyl součástí studie.

5.5 Závěr

Z výše popsaných výsledků vyplývají dvě jasná doporučení pro forenzní praxi: za prvé, pokud porovnáváme krátkodobý spektrální sklon u různých

mluvčích, je nutné zajistit srovnatelné zastoupení jednotlivých vokálů v porovnávaných vzorcích. A naopak – máme-li k dispozici dostatečný počet vokálů ve srovnatelném zastoupení, mohou být ukazatele krátkodobého spektrálního sklonu velmi efektivní pro identifikaci. Zároveň je dobré využít více ukazatelů a neomezit se jen na jediný, protože různé způsoby měření spektrálního sklonu jsou citlivé na různé části spektra a mohou odhalit další odlišnosti.

Spektrální sklon, ať už měřený dlouhodobě (LTAS) nebo krátkodobě, se tedy jeví jako velmi užitečný akustický parametr pro identifikaci mluvčího, jelikož charakteristické vlastnosti jednotlivých hlasů se promítají do tvaru spektra. Je ovšem třeba, abychom si byli vědomi omezení, která s sebou nese analýza spektra a následná kvantifikace sklonu jako jediného čísla. Také musíme brát v úvahu jak lingvistické, tak i nelingvistické faktory ovlivňující spektrum, zejména kvalitu nahrávky, ale také její kvantitu. V tomto smyslu má dlouhodobé spektrum oproti krátkodobému tu nevýhodu, že pro reprezentativní hodnoty parametrů, které ho kvantifikují, vyžaduje delší signál (tradičně se uvádí 30 vteřin), což ne vždy bývá v praxi k dispozici.