Egy egyszerű módszer modális beszéd glottalizálttá alakítására

Bőhm Tamás, Németh Géza

BME Távközlési és Médiainformatikai Tanszék, 1117 Budapest, Magyar Tudósok krt. 2.

Kivonat: A beszédtechnológia számos területén rendkívül hasznos lenne az irreguláris hangszalagrezgés (glottalizáció) megfelelő kezelése. Cikkünk ilyen irányú munkánk első eredményeit ismerteti: egy félautomatikus eljárást mutatunk be, amely képes modális (nem glottalizált) beszédet glottalizálttá alakítani. A korábbi algoritmusok a jitter növelésével próbálták mesterségesen érdes hangzásúvá tenni a beszédet. Módszerünk ezzel szemben alapperiódusok kitörlésével éri el ezt a hatást, amit a megmaradó periódusok amplitúdójának perturbálásával tovább erősít. Formális meghallgatásos tesztben kimutattuk, hogy az így előállított beszédjel természetes hangzású és hasonlóan érdes, mint a természetes glottalizált felvételek.

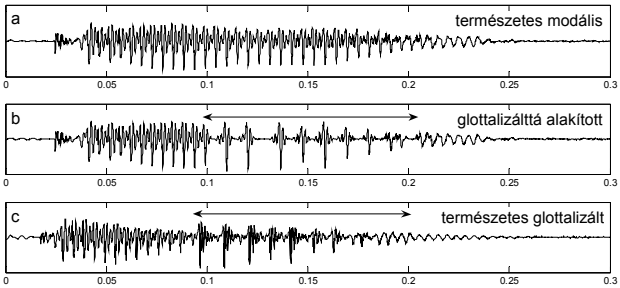
1 Bevezetés

A zöngeképzés, a fonáció során a hangszalagok általában közelperiodikusan (kváziperiodikusan) rezegnek. Ilyenkor a hangszalagok nagyjából állandó időközönként összecsapódnak – a rezgés reguláris, vagy más néven modális. Rövidebb-hosszabb ideig azonban ez a rezgés irregulárissá válhat, azaz hirtelen változások jelentkezhetnek a rezgés pillanatnyi frekvenciájában, amplitúdójában vagy mindkettőben. Ezt a jelenséget nevezzük *glottalizációnak* (1.*c)* ábra), ami gyakori jelenség mind egészséges, mind sérült gégefunkcióval rendelkező beszélők esetén. Gyakran rendkívül alacsony alapfrekvencia és a glottális impulzusok gyors lecsengése kíséri. A glottalizációt érdes, rekedt hangként érzékeljük. A jelenség produkciós és percepciós hátterének áttekintése [1]-ben olvasható.

A glottalizáció előfordulása függ a prozódiai szerkezettől (például gyakran egybeesik prozódiai egységhatárokkal és hangsúlyos szótagokkal [2]) és információt hordoz a beszélő személyéről, nyelvjárásáról [4], hangulatáról és érzelmi állapotáról [3]. Így a glottalizáció megfelelő manipulációja hozzájárulhat természetes hangzású, érzelmi töltettel rendelkező és személyre szabott beszédszintézis rendszerek építéséhez.

Azonban még nincs általánosan alkalmazható módszer modális beszéd glottalizálttá alakítására és fordítva. Számos kísérlet történt glottalizáció előállítására formánsszintézissel (pl. [9]), de ezek több tucat szintézisparaméter kézi beállítását igénylik. A másik módszer egy természetes beszédfelvételen a glottális impulzusok időzítésének perturbálása, azaz a jitter növelése [7,10]. Általánosan elfogadott tény, hogy a jitter összefügg a beszéd érdességével, de sokszor az összefüggés csak áttételes [6] és a különböző típusú perturbációk máshogy befolyásolják az érzeti zöngemi-

nőséget [5]. Ezért ebben a tanulmányban más megközelítést választottunk: úgy próbáljuk a glottalizációra jellemző hullámformát előállítani, hogy egyes kézzel kiválasztott alapperiódusokat kitörlünk a jelből, más periódusokat pedig felerősítünk vagy csillapítunk. Bár módszerünk jelentősen növeli a jelben a jittert, a glottalizáció számos más akusztikai jellegzetességét is reprodukálja.



1. ábra. Egy női bemondótól származó természetes modális (*a*) és glottalizált (*c*) felvétel hullámformája, valamint a modálisból mesterségesen glottalizálttá alakított változat (*b*). A nyilak a glottalizált részeket jelölik.

2 A módszer leírása

Módszerünk hasonlít a PSOLA algoritmusra [8]. Az alkalmazott analízis és szintézis megegyezik azzal, a különbség a manipulációban van: míg a PSOLA az alapfrekvencia módosítása érdekében időben elcsúsztatja az alapperiódusokat, esetünkben ehelyett az egyes periódusok amplitúdóját változtatjuk.

Analízis. A módszer bemenete a beszédjel és a glottális impulzusok időpontjai, azaz a *pitchmark*-ok. Az analízis célja, hogy a jelet szétbontsa alapperiódusokra. Ezt a megfelelő pitchmark környezetének kiablakozásával éri el. Egy olyan Hanning ablakkal szorozza be a beszédjelet, aminek a csúcsa az aktuális pitchmark-on van és az előzőtől a következő pitchmark-ig tart (tehát két alapperiódust fed le).

Manipuláció. Minden egyes kiablakozott alapperiódust beszorzunk egy kézzel beállított s faktorral. Így a periódusokat egyenként felerősíthetjük ($s > 1$), csillapíthatjuk ($s < 1$), kitörölhetjük ($s = 0$) vagy akár módosítás nélkül meghagyhatjuk ($s = 1$). Egyegy periódus kitörlésével érdes hangzás érhető el. A hatás több egymás utáni periódus törlésével és az amplitúdók perturbálásával fokozható. A cél a természetes glottalizációra hasonlító hosszú és irreguláris alapperiódusok létrehozása.

Szintézis. A módosított beszédjelet a faktorokkal megszorzott periódusok átfedve összeadásával (overlap-and-add) kapjuk meg. Ha nem végeztünk semmilyen manipulációt, akkor a kerekítési hibától eltekintve visszakapjuk az eredeti jelet.

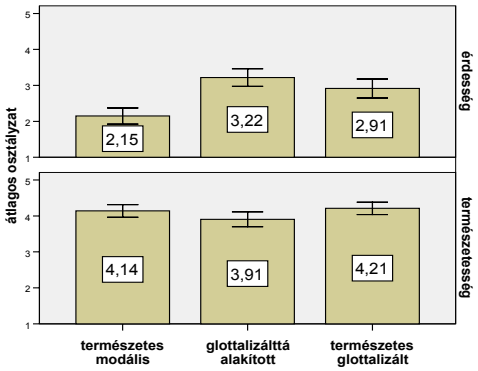
Az 1. ábra *b)* részén az *a)* részen látható felvétel glottalizálttá alakított változata látható. Összehasonlításképpen az ábra *c)* része egy természetes glottalizált felvételt ábrázol, ugyanannak a bemondónak az ejtésében.

3 Értékelés

Egy meghallgatásos kísérlettel értékeltük az eljárást. Négy rövid szót vettünk fel, amit két bemondó modálisan és glottalizált befejezéssel is felolvasott. Módszerünkkel a modális szavak végét glottalizálttá alakítottuk. Megpróbáltuk a bemondó természetes glottalizált kiejtésének glottális impulzus-időzítéseit és -amplitúdóit reprodukálni. A manipulált felvételt meghallgatva iteratívan finomítottuk a beállított *s* faktorokat.

A 12 kísérleti személy két külön tesztet végzett el: az egyikben az érdesség, a másikban a természetesség szempontjából kellett értékelniük a felvételeket. Egyesével, véletlenszerű sorrendben hallották a természetes modális, a természetes glottalizált és a mesterségesen glottalizálttá alakított hanganyagokat[1]. Ezeket egy ötpontos skálán kellett osztályozniuk (1: nagyon természetellenes/egyáltalán nem érdes; 5: nagyon természetes/nagyon érdes). A kísérlet előtt végighallgatták az összes hanganyagot, valamint hallottak pár nagyon érdes és egyáltalán nem érdes példát.

Négy kísérleti személy a természetes glottalizált felvételeket *kevésbé* érdesnek ítélte meg, mint a természetes modálisokat, így osztályzataikat a továbbiakban nem elemeztük. A maradék nyolc személy esetén egyutas varianciaanalízist (ANOVA) végeztünk a hanganyag típusának függvényében, külön a természetesség és külön az érdesség osztályzatokra. Az egyes felvételtípusokhoz tartozó átlagos osztályzatok a 2. ábrán láthatóak. Az alábbiakban részletezett különbségek Tukey-féle post hoc tesztek alapján 5%-os szinten szignifikánsak.



2. ábra. A meghallgatásos kiértékelés során kapott átlagos osztályzatok. A függőleges szakaszok az átlagokhoz tartozó 95%-os konfidencia-intervallumokat jelölik.

Ahogy várható volt, a természetes glottalizált felvételek érdesebb hangzásúak, mint a természetes modálisok. A természetes modális felvételeket glottalizálttá alakítva 1,07 osztályzattal nőtt az érdesség, ami így már nem tér el szignifikánsan a célértéktől (a természetes glottalizált hanganyagok érdességétől). Az átalakítás azonban

[1] A kísérleti személyek más típusú (pl. formánsszintetikus) hanganyagokat is értékeltek, amelyek egy másik tanulmány részét képezték, itt irrelevánsak.

csak elhanyagolható, nem szignifikáns romlást okozott a természetesség megítélésében.

4 Összefoglalás

Egy olyan egyszerű, félautomatikus, pitch-szinkron beszédfeldolgozási eljárást mutattunk be, amely képes modális beszédet glottalizálttá alakítani egyes alapperiódusok kinullázásával, valamint más periódusok amplitúdójának megváltoztatásával. A meghallgatásos értékelés azt mutatta, hogy a transzformált felvételeket a hallgatók hasonlóan érdesnek és természetesnek ítélték meg, mint a természetesen glottalizáltakat. Mivel az alkalmazott analízis és a szintézis módszerek megegyeznek a PSOLA algoritmus megfelelő feldolgozási szakaszaival, módszerünk könnyen integrálható PSOLÁ-t alkalmazó beszédszintetizátorokba. Ehhez azonban az eljárás automatizálása, azaz az *s* faktorok algoritmikus beállítása, valamint az ellentétes irányú transzformáció megvalósítása is szükséges. A szerzők jelenleg ebben a két irányban folytatják a munkát.

Bibliográfia

1. Blomgren, M., Chen, Y., Ng, M.L., Gilbert, H.R.: Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. JASA 103 (1998) 2649–2658
2. Dilley, L., Shattuck-Hufnagel, S., Ostendorf, M.: Glottalization of word-initial vowels as a function of prosodic structure. J. Phonetics 24 (1996) 423–444
3. Gobl, C., Ní Chasaide, A.: The role of voice quality in communicating emotion, mood and attitude. Sp. Comm. 40 (2003) 189–212
4. Henton, C.G., Bladon, A.: Creak as a sociophonetic marker. In: Hyman, L.M., Li, C.N. (eds.): Language, speech and mind. Routledge, London (1987) 3–29
5. Hillenbrand, J.: Perception of aperiodicities in synthetically generated voices. JASA 83 (1988) 2361–2371
6. Kreiman, J., Gerratt, B.R.: Perception of aperiodicity in pathological voice. JASA 117 (2005) 2201–2211
7. McCree, A.V., Barnwell, T.P.: A mixed excitation LPC vocoder model for low bit rate speech coding. IEEE Trans. Speech and Audio Proc. 3 (1995) 242–249
8. Moulines, E., Charpentier, F.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Sp. Comm. 9 (1990) 453–467
9. Pierrehumbert, J.B., Frisch, S.: Synthesizing Allophonic Glottalization. In: van Santen, J.P.H. et al. (eds.): Progress in Speech Synthesis. Springer, New York (1997) 9–26
10. Verma, A., Kumar, A.: Introducing rougness in individuality transformation through jitter modeling and modification. Proc. ICASSP (2005) 5–8

A szerzők szeretnék köszönetüket kifejezni Stefanie Shattuck-Hufnagelnek szakmai útmutatásáért és a Fulbright Bizottságnak a kísérleti személyek toborzásáért. A kutatást részben az NKFP 2. keretprogramja (szerződésszám: 2/034/2004) támogatta.