Magdalena IGRAS, Wiesław WSZOŁEK
AGH AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE
{migras@agh.edu.pl, }

Pomiary parametrów akustycznych mowy emocjonalnej – krok ku modelowaniu wokalnej ekspresji emocji

Streszczenie

Niniejsza praca podejmuje próbę analizy emocji podstawowych w mowie i stanowi jednocześnie podłoże do dalszych badań na tym polu. Zaprojektowano i wykonano korpus mowy emocjonalnej oraz przeprowadzono wstępne testy mające na celu ocenę jego przydatności do automatyzacji wykrywania emocji w mowie. Na etapie etykietowania nagrań przeprowadzono testy percepcyjne badające subiektywną klasyfikację emocji w nagraniach zarówno z treścią, jak i pozbawionych treści. Przetwarzanie nagrań pozwoliło na wyodrębnienie szeregu cech charakterystycznych dla emocji podstawowych, co pozwoliło opracować robocze wokalne profile emocjonalne będące zestawieniem cech akustycznych cech różnicujących poszczególne stany emocjonalne. W oparciu o wektory powyższych cech przetestowano metodę automatycznej klasyfikacji emocji przy użyciu metody kNN oraz samoorganizującej się sieci neuronowej. Artykuł prezentuje również propozycje aplikacji medycznych opartych na pomiarach emocji w głosie.

Abstract

The paper presents an approach to analysis of emotional content of speech signal and constitutes a base for further research in this field. A corpus of emotional speech was designed and recorded. Preliminary tests were performed to evaluate its applicability to automatic emotion recognition. On the stage of recordings labeling, human perceptual tests were performed (using recordings with and without semantic content). Further signal processing allowed to extract a set of features characteristic for each emotion, and led to developing preliminary vocal emotions profiles (sets of acoustic features characteristic for each of basic emotions). Using selected features vectors, methods of automatic classification (kNN and self organizing neural network) were tested. The article presents also a discussion of using the results of this kind of research for medical applications.

Słowa kluczowe: rozpoznawanie emocji, wokalne korelaty emocji, sygnał mowy

Keywords: emotions recognition, vocal correlates of emotions, speech signal

Title: Emotional speech acoustic parameters measurement – a step towards vocal emotions expression modelling

1. Wprowadzenie

Stany afektywno-kognitywne człowieka pełnią kluczową rolę w interakcjach międzyludzkich. Jako że głos to jeden z naturalnych, spontanicznych środków ekspresji emocji, sygnał mowy może posłużyć do skutecznej detekcji i identyfikacji stanów emocjonalnych mówcy.

Problematyka badania emocji w głosie jest zagadnieniem interdyscyplinarnym, które angażuje nauki humanistyczne, medyczne i techniczne. Wraz z rozwojem technologii mowy (systemy automatycznego rozpoznawania mowy i automatycznego rozpoznawania mówcy) podejmowane są próby opracowania systemów automatycznie rozpoznających emocje w głosie mówcy. Pierwszym krokiem ku temu jest opracowanie technicznego opisu cech sygnału akustycznego znamiennych dla ekspresji poszczególnych emocji. Pośród licznych opracowań dla innych języków [1][6], badania takie w polskiej nauce wciąż należą do rzadkości [3].

2. Materiał akustyczny

Aby uzyskać nagrania sygnału mowy emocjonalnej przy zachowaniu poprawności etycznej, odpowiedniej jakości sygnału, a zarazem najbardziej korzystnej struktury nagrań (gdzie jedyną zmienną będzie stan emocjonalny), zaprojektowano korpus Emotive. Do nagrań użyto mikrofonu pojemnościowego AKG C5 Vocal oraz rejestratora Zoom H4N, uzyskując nagrania w formacie .wav 16 bit o częstotliwości próbkowania 44 100 Hz i poziomie SNR > 30 dB. Dokonano rejestracji nagrań mowy emocjonalnej kilkunastu mówców – aktorów, studentów aktorstwa, osób o przygotowaniu teatralnym. Dla każdej osoby zarejestrowano nagrania o tej samej treści (pojedyncze słowa, zdania dialogowe, tekst ciągły) w każdym z podstawowych stanów emocjonalnych (radość, smutek, strach, złość, zdziwienie, stan neutralny). Uzyskano dzięki temu szereg usystematyzowanych nagrań etykietowanych intencją mówcy.

Nagrania poddano selekcji w oparciu o wyniki testów percepcyjnych, w których grupa słuchaczy oceniła każde z nich pod kątem zawartości emocjonalnej nagrania. Ich ocena posłużyła do ponownego etykietowania nagrań oceną słuchaczy. Do dalszego przetwarzania zostały wybrane nagrania o tożsamych etykietach: intencjonalnej i percepcyjnej.

Zamieszczona poniżej tabela prezentuje procentowo klasyfikację dokonaną przez słuchaczy (w kolumnach – nagrania etykietowane intencją mówcy, w wierszach – ocena słuchaczy).

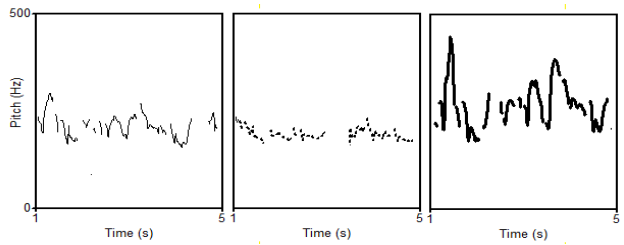
	ne	*ra*	*sm*	*st*	*zd*	*zl*
ne	**59**	7	10	1	2	0
ra	3	**42**	0	0	7	1
sm	7	0	**78**	13	0	3
st	1	0	3	**38**	4	3
zd	7	4	0	10	**75**	0
zl	1	2	0	3	4	**76**
nie	21	46	9	35	9	17

Tab.1. Tablica błędu dla testów percepcyjnych (wartości procentowe). Oznaczenia: ne – stan neutralny, ra – radość, sm – smutek, st – strach, zd – zdziwienie, zl – złość, nie – nie rozpoznano.

3. Przetwarzanie sygnałów

3.1. Parametryzacja i ekstrakcja cech

Wyselekcjonowane nagrania poddano wstępnej normalizacji i preemfazie. Następnie dokonano parametryzacji sygnału oraz ekstrakcji szeregu cech sygnału, znormalizowanych dla każdej emocji do stanu neutralnego dla każdego mówcy jako referencji. [2][4][5]



Rys. 1 Przykład przebiegów konturów intonacji (F0) fragmentu wypowiedzi tego samego mówcy o tej samej treści w stanach: neutralny (linia ciągła), smutek (linia przerywana), złość (linia ciągła pogrubiona).

3.2. Klasyfikacja

Przeprowadzono testy klasyfikatorów: kNN oraz samoorganizującej się sieci neuronowej, używając programu Matlab. [3]

3.3. Wokalne profile emocjonalne

Jako wynik analizy zbiorczej cech różnicujących emocje, sporządzono robocze wokalne profile wybranych emocji podstawowych. Kluczowe okazały się cechy prozodyczne sygnału związane z częstotliwością i energią (średnia F0, zakres F0, charakterystyka zmian F0, jitter, shimmer, moc sygnału, rozkład energii) oraz tempo mówienia (czas trwania wypowiedzi, ilość i czas pauz). [1]

Parametr	*Smutek*	*Stan neutralny*	*Radość*
F0 średnia	F0sr$^{\downarrow}$	F0sr	F0sr$^{\uparrow}$
F0 – odchylenie standardowe	F0od$^{\downarrow}$	F0od	F0od$^{\uparrow}$
Zakres F0	F0za$^{\downarrow}$	F0za	F0za$^{\uparrow}$
Energia	E$^{\downarrow}$	E	E$^{\uparrow}$

Tab.2. Fragment zestawienia względnych zmian cech akustycznych wybranych emocji ($^{\downarrow}$ - spadek; $^{\uparrow}$- wzrost)

4. Wnioski

Badania nad automatyczną detekcją emocji w sygnale mowy opierają się na założeniu, że istnieją uniwersalne wzorce wokalnego komunikowania poszczególnych emocji. [1] W praktyce bardzo dużym utrudnieniem jest subiektywność i różnice indywidualne zarówno w ekspresji jak i percepcji emocji w mowie, uwarunkowane różnicami osobniczymi. Na etapie etykietowania nagrań ocena pewnej grupy statystycznej nie zapewnia obiektywności – poziom odbioru i oceny emocji zależy od indywidualnej wrażliwości, empatii i inteligencji emocjonalnej. Dodatkowo jednoznaczna klasyfikacja była często niemożliwa, ponieważ najczęściej występują interkorelacje i współwystępowanie złożonych stanów emocjonalnych.

Dywersyfikacja indywidualnego sposobu ekspresji i percepcji emocji powoduje nieodzowną potrzebę kalibrowania systemu automatycznej detekcji emocji pod kątem charakterystyki profilu danego użytkownika afektywnego interfejsu głosowego. [1][3]

5. Zastosowania

Badania nad rozpoznawaniem emocji w mowie mają znaczenie zarówno poznawcze, jak i wdrożeniowe. Obok całego spektrum zastosowań technologicznych (m.in. jako moduł systemów automatycznego rozpoznawania mowy i mówcy), aplikacje zbudowane w oparciu o algorytmy identyfikujące stan emocjonalny pacjenta w oparciu o parametry akustyczne jego mowy mają duży potencjał w dziedzinie diagnostyki medycznej. W psychologii i psychiatrii pomiary parametrów głosu pacjentów mogą służyć jako metoda diagnostyczna (np. depresji, choroby afektywnej dwubiegunowej, ADHD).[3] Periodyczne badanie zabarwienia emocjonalnego głosu może służyć jako jedno z narzędzi wspomagających monitoring przebiegu leczenia zaburzeń psychologicznych i neurologicznych (np. zwiększenie tempa wypowiedzi i skrócenie długości pauz może świadczyć o sukcesie terapeutycznym i wychodzeniu pacjenta z choroby). Kolejnym potencjalnym zastosowaniem jest wczesne diagnozowanie stresu stanowiącego podłoże wielu chorób cywilizacyjnych.

6. Literatura

[1] Izdebski K.: Emotions in the human voice, Vol. I-III, Plural Publishing, San Diego 2008.

[2] Tadeusiewicz R.: Sygnał mowy, WKiŁ, Warszawa 1987.

[3] Ciota Z.: Metody przetwarzania sygnałów akustycznych w komputerowej analizie mowy, wyd. EXIT Warszawa 2010.

[4] Ziółko M., Ziółko B.: Przetwarzanie mowy; Wydawnictwa AGH, Kraków 2011.

[5] Boersma, Paul: Praat, a system for doing phonetics by computer. Glot International 5:9/10, 341-345.

[6] Waaramaa-Maki-Kulmala T.: Emotions in voice - acoustic and perceptual analysis of voice quality in the vocal expression of emotions, University of Tampere 2009

Praca naukowa finansowana ze środków na naukę jako projekt badawczy POIG INSIGMA 01.01.02-00-062/09