STUDIA INFORMATICA 2013
Volume 34 Number 2B (112)

Magdalena IGRAS, Bartosz ZIÓŁKO
AGH Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie,
Katedra Elektroniki

BAZA DANYCH NAGRAŃ MOWY EMOCJONALNEJ

Streszczenie. Artykuł prezentuje opracowaną w AGH bazę danych nagrań mowy emocjonalnej, zgromadzoną w celu badań nad zawartością afektywną sygnału mowy. Opisano sposób rejestracji, parametry, strukturę, metadane i licencję bazy danych. Przedstawiono przykładowe zastosowania do opracowania metod detekcji stanów emocjonalnych w głosie oraz normalizacji nagrań na potrzeby ASR.

Słowa kluczowe: nagrania mowy, detekcja emocji w głosie

DATABASE OF EMOTIONAL SPEECH RECORDINGS

Summary. The paper presents a database of emotional speech recordings collected in AGH for research on affective content of speech signal. We describe the method of data acquisition, the parameters, structure, metadata and license. We present example applications for development of the methods of emotions detection in voice and emotional speech normalization for ASR.

Keywords: speech recordings, detection of emotions in speech

1. Wprowadzenie

Poniższe opracowanie zawiera informacje o nowopowstałej bazie nagrań mowy emocjonalnej w języku polskim. W rozdziale 1 opisano kontekst zapotrzebowania na tego typu bazy oraz dokonano przeglądu istniejących baz. Na ich tle zaprezentowano nasz korpus, podając w rozdziale 2 jego szczegółowy opis, wraz z parametrami technicznymi nagrań. Rozdział 3 poświęcono podsumowaniu doświadczeń i uwag z procesu tworzenia bazy. W rozdziale 4 nakreślono potencjalne zastosowania zebranych danych.

1.1. Znaczenie baz mowy emocjonalnej dla technologii mowy

Rozwój interfejsów komunikacji człowieka z komputerem podąża w stronę zapewnienia całkowicie nieabsorbujących metod interakcji. Wykorzystywane dotychczas urządzenia, czyli klawiatury, myszy i ekrany, są wspomagane mikrofonami, kamerami i ekranami dotykowymi. Celem tego jest opracowanie interfejsów pozwalających na całkowicie naturalną i intuicyjną komunikację z wykorzystaniem ludzkich zmysłów, tj. wzroku, słuchu, dotyku oraz mowy. Jako że mowa jest najbardziej naturalnym sposobem porozumiewania się, wraz z dynamicznym wzrostem jakości i szybkości przetwarzania danych oraz rozwojem społeczeństwa informacyjnego zauważalny jest wyraźny trend używania sygnału mowy jako sposobu interakcji w interfejsach człowiek-komputer [17].

Z punktu widzenia technologii mowy w sygnale mowy zawarte są informacje:

- semantyczne (dotyczące treści wypowiedzi – komunikacja werbalna),
- osobnicze (stanowiące o unikalności głosu każdego człowieka, pozwalające na identyfikację/weryfikację mówcy),
- emocjonalne (pozwalające na określenie emocji osoby mówiącej – komunikacja ekspresywna).

Emocje wyrażane w mowie stanowią ważną część komunikatu, uzupełniającą komunikat słowny. Wsparcie systemów rozpoznawania mowy możliwościami detekcji emocji mówcy ma szansę poprawić wygodę korzystania z interfejsów głosowych. Oznacza to uzyskanie nowej jakości działania systemów interakcji głosowej, które będą adaptowane do nastawienia danego użytkownika. Docelowo poprawi to jakość aplikacji i zadowolenie użytkowników – umożliwi lepsze zrozumienie wpływu cech głosu ludzkiego na efektywność rozpoznawania mowy. Waga zagadnienia jest o tyle fundamentalna, że wiedza o tym, jakie cechy sprawiają, że ludzki głos jest nośnikiem informacji o jego emocjach, dotyczy codziennych zjawisk komunikowania się ludzi i funkcjonowania w społeczeństwie.

Rozwój systemów technologii mowy, realizujących detekcję emocji, wymaga odpowiednich zasobów nagrań. Jednym z największych wyzwań stojących przed analizą mowy emocjonalnej jest posiadanie odpowiednio dużej i różnorodnej pod względem treści oraz mówców bazy nagrań mowy emocjonalnej [8]. Niektórzy specjaliści uważają nawet, że jest to jeden z najważniejszych elementów warunkujących tempo rozwoju tej młodej dziedziny nauki [5]. O istotności zagadnienia mogą świadczyć liczne monograficzne publikacje [2, 3, 4], poświęcone wyłącznie problematyce gromadzenia materiału eksperymentalnego.

Detekcja emocji podstawowych (radość, złość, smutek, zdziwienie, wstręt, strach) [11-13] na podstawie sygnału mowy opiera się na paradygmacie Scherera [12, 13], zakładającym, że emocje podstawowe mogą być scharakteryzowane w sposób uniwersalny przez jednoznaczny

wzorzec lub konfigurację parametrów akustycznych, stąd odpowiednio duży i zróżnicowany zbiór nagrań stanowi bazę wzorców emocji w mowie.

1.2. Przegląd istniejących korpusów

Pożądane parametry korpusów mowy emocjonalnej to: autentyczność emocji, różnorodność emocji, jednoznaczność emocji, dobra jakość nagrań, jak największa liczba mówców, jednolita treść. W praktyce wszystkie te cechy są niemożliwe do uzyskania jednocześnie. Wady i zalety wyboru poszczególnych rozwiązań zestawiono w tabeli 1.

Biorąc pod uwagę autentyczność emocji i rodzaj etykietowania, Sidorova [1] wyróżnia następujące rodzaje korpusów mowy emocjonalnej:

- Emocji odgrywane, uzyskiwane przez nagrywanie aktorów poproszonych o odgrywanie predefiniowanych emocji, identyfikowanych następnie przez słuchaczy.
- Emocje autentyczne, identyfikowane następnie przez słuchaczy. Jako źródło nagrań spontanicznych emocji używane są relacje na żywo z wydarzeń sportowych oraz programy typu talk-show [11].
- Emocje wywoływane sztucznie, etykietowane przez mówców. Jako najczęściej stosowane techniki indukowania emocji stosuje się emocjonalne filmy, opowieści, obrazy, wyobrażenia, gry komputerowe [12].

Tabela 1

Porównanie cech wybranych rodzajów korpusów mowy emocjonalnej

	Wady	Zalety
Emocje indukowane	Restrykcje etyczne, Zmienność osobnicza, Niska intensywność, Sztuczność sytuacji eksperymentalnej	Dobra jakość nagrań, Spontaniczność reakcji
Baza nagrań z mediów	Niejednolita i najczęściej słaba jakość nagrań, szumy, artefakty, nakładające się głosy	Różnorodność sytuacji, Najwyższa wiarygodność, Autentyczność – emocje indukowane rzeczywistymi bodźcami
Emocje odgrywane	Sztuczność emocji	Dobra jakość nagrania Możliwość uzyskania gamy emocji przy dowolnej treści, Dokładne zaplanowanie
Bazy nagrań z telefonu alarmowego	Niejednolita i najczęściej słaba jakość nagrań	Autentyczność emocji

Ververidis i Kotropoulos [2] dokonali zestawienia 32 korpusów mowy emocjonalnej (w tym 11 w języku angielskim, 8 – niemieckim, 3 – japońskim, 2 – holenderskim, 2 – hiszpańskim, 1 – duńskim, 1 – hebrajskim, 1 – szwedzkim, 1 – chińskim, 1 – rosyjskim oraz 1

wielojęzyczna). Ponad 20 z nich zawiera nagrania emocji symulowanych. Część z nich jest publicznie dostępna.

Pośród polskich zasobów korpusów mowy emocjonalnej jest znacznie mniej. W ramach prac w Instytucie Telekomunikacji Politechniki Warszawskiej stworzono dla języka polskiego bazę emocji spontanicznych BES, zawierającą 370 nagrań o różnym nacechowaniu emocjonalnym, pochodzących z audycji Polskiego Radia i TVP (Janicki [6]). Z kolei Demenko i inni [7] stworzyli bazę nagrań z telefonu alarmowego – spośród 60 000 połączeń telefonicznych z numerem alarmowych 997 Policji w Poznaniu automatycznie wyselekcjonowano 22 000 dialogów, z których kilkaset wybrano do analizy akustycznej (ostatecznej selekcji dokonano na podstawie oceny percepcyjnej). Cichosz [8] na potrzeby swoich badań wykonał nagrania emocji odgrywanych przez aktorów. Było to 5 zdań wypowiadanych w 6 stanach emocjonalnych (złość, smutek, strach, radość, nuda, neutralny) przez 8 mówców.

W innych polskich pracach [9, 10] użyto własnych nagrań, wykonanych na potrzeby badań, lub nagrań pochodzących z mediów, zgromadzonych na potrzeby badań. Niestety, wg najlepszej wiedzy autorów nie ma żadnych informacji o publicznej dostępności wymienionych zasobów.

2. Opis bazy nagrań

Powstały w AGH korpus nagrań jest jedynym w Polsce usystematyzowanym zbiorem nagrań emocji odgrywanych, dostępnym w ramach licencji.

2.1. Rodzaj emocji

Korpus zawiera nagrania wyrażające pięć spośród emocji podstawowych (radość, smutek, złość, strach, zdziwienie), ironię oraz stan neutralny/obojętny, jako sygnał referencyjny. Ironia /sarkazm/drwina nie są emocjami w rozumieniu teorii emocji podstawowych. Można uznawać je za emocje złożone, postawę emocjonalną bądź środek wyrazu, świadczący o nastawieniu emocjonalnym. Ton ironiczny jest dodatkową informacją niesioną przez sygnał mowy – reprezentującą postawę emocjonalną wobec wypowiadanej treści.

2.2. Mówcy i treść

W nagraniach wzięło udział 12 mówców (6 kobiet, 6 mężczyzn) w wieku 20-30 lat. Część z nich to profesjonalni aktorzy lub amatorzy, a część – studenci wolontariusze.

Tabela 2

Zawartość słowna korpusu nagrań

ZDANIA

1. Dzień dobry.
2. Witam serdecznie.
3. Miło cię widzieć.
4. Dlaczego nie?
5. Muszę z tobą porozmawiać.
6. OK. Do zobaczenia!
7. Pomóż mi.
8. Wzajemnie.
9. Nie chcę.
10. U mnie super.
11. Świetny pomysł!
12. Chyba Ci się nudzi.
13. Bardzo się cieszę.
14. On jest najlepszy na świecie.
15. Czy naprawdę tak myślisz?
16. Która jest teraz godzina?
17. Przecież ty nic nie wiesz.
18. To był żart.
19. No oczywiście.
20. Dziękuję, rozumiem.
21. Mam dość.
22. Skąd mnie znasz?
23. Myślę, że tak.
24. Masz rację.
25. Na pewno tak jest.
26. Dziś jest bardzo ładna pogoda.
27. To bardzo dobrze.
28. Proszę bardzo.
29. Nic nie rozumiesz.
30. Mam rację.
31. Przepraszam cię.
32. Ależ skąd!
33. A jednak.
34. Dziękuję za rozmowę.
35. Wcale nie trzeba.
36. Nie, dziękuję.
37. Nie wiem.
38. Nie trzeba.
39. Nie denerwuj się tak.
40. Pokaż, co potrafisz.
41. O czym Ty mówisz?
42. Dziękuję za pomoc.
43. Nieźle.
44. Uśmiechnij się.
45. Idę do domu.
46. Miłego dnia!

POLECENIA

1. Nowy	7. Lewo
2. Otwórz	8. Góra
3. Usuń	9. Dół
4. Cofnij	10. Start
5. Zapisz	11. Stop
6. Prawo	12. OK

CYFRY

1. Jeden	6. Sześć
2. Dwa	7. Siedem
3. Trzy	8. Osiem
4. Cztery	9. Dziewięć
5. Pięć	0. Zero

TEKST

Gabinet zebrał się na oczekiwanym z wielkim zainteresowaniem nadzwyczajnym posiedzeniu. Równolegle do zabiegów premiera, prezydent prowadzi nieoficjalne konsultacje w sprawie wyjścia z sytuacji patowej, w jakiej znalazły się nie tylko gospodarka i finanse, ale również polityka.

Nie będzie zapowiadanych wcześniej inicjatyw, jak jednorazowe opodatkowanie wszystkich kont bankowych bez względu na ich stan, ani przywrócenia podatku od nieruchomości, nieobowiązującego mieszkań, w których zameldowany jest sam właściciel.

Po uprzednim zapoznaniu się z tekstami, mówcy zostali poproszeni o przeczytanie ich kolejno, w sposób wyrażający poszczególne emocje. Dla każdego mówcy zarejestrowano nagrania o tej samej treści (około 10 minut dla każdego mówcy). Treść nagrań stanowią poje-

dyncze słowa (cyfry, polecenia sterujące), zdania dialogowe (46 zdań z naturalnych codziennych rozmów) oraz jeden tekst ciągły (tab. 2). Treść dobrano tak, żeby była możliwie neutralna i nie indukowała konkretnej emocji. Treść zaprojektowano pod kątem użycia w interfejsach człowiek-komputer oraz różnorodności typów wypowiedzi. Łącznie dla każdego mówcy zarejestrowano 282 słowa, każde w 6 stanach emocjonalnych.

2.3. Parametry audio

Wypowiedzi nagrywano za pomocą rejestratora Zoom H4N oraz mikrofonu pojemnościowego AKG C5 Vocal i dynamicznego AKG Shotgun C568. Uzyskane nagrania mają postać plików PCM .wav, o parametrach: częstotliwość próbkowania 44 100 Hz, rozdzielczość 16 bit, SNR średnio ok. 40 dB.

2.4. Struktura i rozmiar bazy

Nagranie każdego mówcy zostało podzielone na części tematyczne (osobno: zdania, tekst ciągły, cyfry, polecenia), znajdujące się w osobnych plikach. Łączna wielkość zarchiwizowanych danych (całego korpusu) to 1,5 GB.

2.5. Metadane

W warstwie metadanych nagrania oznaczone są akronimem mówcy, informacją, czy mówca jest aktorem oraz nazwą emocji.

2.6. Aspekty prawne

Nagrania mają uregulowaną sytuację prawną, umożliwiającą ich przetwarzanie (zgoda mówców na wykorzystanie naukowe oraz przetwarzanie nagrań w systemach informatycznych technologii mowy, w tym komercyjnych) oraz prezentowanie publiczne (zgoda na anonimowe odtwarzanie na konferencjach, wykładach i prezentacjach systemów technologii mowy).

Możliwe jest nabycie licencji na wykorzystanie prezentowanej bazy nagrań do celów naukowych lub komercyjnych.

3. Doświadczenia i dyskusja

Jak zasygnalizowano w podrozdziale 1.2, nie jest możliwe spełnienie jednocześnie wszystkich kryteriów idealnej bazy nagrań emocjonalnych, co decyduje o problematyczności zaplanowania i utworzenia dobrej jakości bazy. Rozwiązanie zawsze jest kompromisem między poszczególnymi parametrami (tab. 1). U źródeł problemów leży multimodalna natura procesów związanych z percepcją i ekspresją emocji. Najczęściej stosowane modele emocji podstawowych są wyidealizowane, podczas gdy w rzeczywistym świecie występują częściej w postaci złożonej – następstwem takich interkorelacji emocji są też trudności w jednoznacznym określaniu emocji w sygnale mowy.

Należy mieć na uwadze wiele zmiennych towarzyszących zjawisku ekspresji emocji w głosie – przede wszystkim różnice indywidualne, determinujące intensywność, a często również formę ekspresji oraz subiektywność odbiorcy oceniającego nagranie – poziom odbioru i oceny emocji zależy od indywidualnej wrażliwości i zdolności do empatii. Cechy emocji mogą też być specyficzne dla języka i kręgu kulturowego/społecznego. Planując wykonywanie nagrań emocji indukowanych bodźcem, należy mieć na względzie aspekty etyczne, bowiem wywoływanie emocji zbyt silnych mogłoby nieść negatywne implikacje dla psychiki osoby badanej, podczas gdy zbyt słabe emocje niedostatecznie odzwierciedlają się w głosie.

Przy tworzeniu niniejszej bazy zdecydowano się na zastosowanie scenariusza emocji odgrywanych. Pozwala on na otrzymanie nagrań o dokładnie zaplanowanej strukturze (ta sama treść dla różnych emocji), nie angażując przy tym czynnika etycznego. Słabą stroną przyjętego rozwiązania jest brak autentyczności (prawdziwej reakcji emocjonalnej, leżącej u podłoża ekspresji emocji w głosie) oraz spontaniczności (zawartość emocjonalna jest wynikiem świadomej intencji, a nie reakcji na bodziec).

4. Opis dalszych prac na bazą

W dotychczasowej formie nagrania etykietowane są intencją mówcy. Baza zostanie uzupełniona o następny rodzaj etykietowania – opinią słuchaczy. Zostaną przeprowadzone testy percepcyjne, polegające na odsłuchaniu przez statystyczną grupę słuchaczy wybranych losowo nagrań (odsłuchanie wszystkich nagrań przez wszystkich słuchaczy jest trudne do zrealizowania ze względu na ich dużą łączną długość, co skutkowałoby zmęczeniem percepcji, schematyzacją reakcji i potencjalnie zafałszowaniem wyników) oraz przypisanie im emocji według ich subiektywnej opinii.

Według Scherera [13], typowe wyniki rozpoznawania, uzyskiwane w tego typu testach, kształtują się na poziomie 60-75% poprawnych klasyfikacji. Taka metoda niesie ze sobą niepewność, będącą wynikiem różnic indywidualnych w percepcji poszczególnych słuchaczy (por. rozdz. 3).

Zostanie stworzona aplikacja udostępniająca dane zawarte w bazie i umożliwiająca szybki dostęp do wybranych zasobów dzięki zastosowaniu filtrów wg kryteriów użytkownika aplikacji. W warstwie metadanych opisy nagrań zostaną wzbogacone o uzyskane w wyniku automatycznego przetwarzania wybrane parametry, będące znamiennymi korelatami emocji w mowie – m.in. wartości średnie i odchylenia częstotliwości podstawowej, energia sygnału, liczba i długość pauz.

Nagrania zostaną również poddane transkrypcji fonetycznej oraz anotacji czasowej na słowa i fonemy, przy użyciu narzędzi do półautomatycznej anotacji w standardzie .mlf [16].

Ponadto, baza będzie uzupełniana o nagrania kolejnych mówców wg tego samego schematu.

5. Zastosowania

Jak podsumowuje Ververidis [2], większość baz mowy emocjonalnej jest tworzona na potrzeby prac nad automatyczną detekcją emocji w mowie, zazwyczaj jako części systemów automatycznego rozpoznawania mowy. Liczne spośród nich są używane również dla potrzeb opracowania algorytmów syntezy mowy emocjonalnej oraz badania ludzkiej percepcji emocji. Niektóre z nich dedykowane są specyficznym zastosowaniom, jak stworzenie oprogramowania typu wirtualny nauczyciel.

Utworzona przez nas baza może służyć zidentyfikowaniu wzorców reakcji emocjonalnych w sygnale mowy i zweryfikowaniu możliwej jakości automatycznego wnioskowania na temat nastawienia emocjonalnego mówcy (cechy wysokopoziomowe mowy) przez algorytmy przetwarzania sygnałów i automatycznej klasyfikacji. Trwają pierwsze tego typu prace na podstawie niniejszej bazy [14, 15].

Praca nad technicznym opisem emocji w mowie ma wielorakie zastosowania. W dziedzinach medycznych może przyczynić się do wspomagania diagnozowania stanu chorych bądź skuteczności leczenia w przypadkach zaburzeń psychologicznych i neurologicznych (autyzm, schizofrenia, ADHD, choroba afektywna dwubiegunowa). Automatyzacja detekcji emocji w mowie może być zastosowana w wielu aplikacjach komercyjnych, jak wspomaganie pracy *call centers* (por. podrozdz. 1.1), tworzenie wirtualnych awatarów, wzbogaconych o syntezę mowy emocjonalnej, oraz wprowadzenie detekcji emocji w różnego rodzaju innych interfejsach głosowych, celem podwyższenia komfortu użytkowania aplikacji i jej efektywności.

Ze względu na strukturę korpusu, może on posłużyć do opracowania algorytmów normalizacji głosu pod kątem zawartości emocjonalnej, na potrzeby systemów automatycznego rozpoznawania mowy.

6. Podsumowanie

W artykule opisano nowo powstałą bazę wzorców emocjonalnych języka mówionego. Korpus zawiera bardzo dobrej jakości nagrania mowy nacechowanej emocjonalnie, pochodzącej od 12 mówców (6 kobiet, 6 mężczyzn), aktorów lub osób o przygotowaniu teatralnym oraz wolontariuszy. Dla każdego mówcy zarejestrowano nagrania o tej samej treści (zdania dialogowe, cyfry, polecenia, tekst ciągły – łącznie 282 słowa), siedmiokrotnie – za każdym razem w jednym z następujących stanów emocjonalnych: radość, smutek, złość, strach, zdziwienie, ironia oraz jako stan referencyjny – stan neutralny. Łączny czas wszystkich nagrań wynosi ponad 3,5 godziny. Nagrania mają postać plików WAV 16 bit, o częstotliwości próbkowania 44 100 Hz. Do nagrań dołączona jest transkrypcja ortograficzna.

Korpus nagrań mowy emocjonalnej jest kluczowym zasobem dla badań nad zawartością emocjonalną sygnału mowy. Zaprezentowany korpus jest jedyną dostępną komercyjnie bazą emocji odgrywanych w języku polskim. Jego zaletą jest struktura odpowiednia dla badań porównawczych – dla każdego mówcy taka sama treść w kilku stanach emocjonalnych.

Zaprezentowano problematykę i znaczenie baz mowy emocjonalnej dla technologii mowy, uwzględniając trudności metodologiczne przy projektowaniu i gromadzeniu takich baz. Dokonano syntezy typów korpusów oraz ich krótkiego przeglądu, ze szczególnym uwzględnieniem baz polskich. Nakreślono również przykładowe zastosowania korpusu, zarówno w dziedzinach medycznych, jak i rozwoju technologii mowy oraz aplikacji komercyjnych.

Podziękowania: Projekt został sfinansowany przez Narodowe Centrum Nauki na podstawie decyzji 2011/03/B/ST7/00442.

BIBLIOGRAFIA

1. Sidorova J.: Speech Emotion Recognition. DEA report, doctoral program Ci`encia Cognitiva i Llenguatge, Universitat Pompeu Fabra, Barcelona 2007.
2. Ververidis D., Kotropoulos C.: A Review of Emotional Speech Databases. Proc. 9th Panhellenic Conference on informatics (PCI), Thessaloniki, Greece 2003, s. 560÷574.

3. Campbell N.: Databases of emotional speech. ISCA Workshop on Speech and Emotion, Belfast 2000, s. 34÷38.
4. Douglas-Cowie E., Campbell N., Cowie R., Roach P.: Emotional speech: towards a new generation of databases. Speech Communication Special Issue Speech and Emotion, No. 40 (1-2), 2003, s. 33÷60.
5. Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., Weiss B.: A Database of German Emotional Speech. Proc. of Interspeech, Lizbon 2005, s. 1517÷1520.
6. Janicki A., Turkot M.: Rozpoznawanie stanu emocjonalnego mówcy z wykorzystaniem maszyny wektorów wspierających (SVM). KSTiT 2008, Bydgoszcz 2008.
7. Demenko G., Jastrzębska M.: Analiza stresu głosowego w rozmowach z telefonu alarmowego. XVIII Konferencja Inżynierii Akustycznej i Biomedycznej, Zakopane 2011.
8. Cichosz J.: Wykorzystanie wybranych cech sygnału mowy do rozpoznawania i modelowania emocji dla języka polskiego. Rozprawa doktorska, Politechnika Łódzka, Łódź 2008.
9. Ciota Z.: Metody przetwarzania sygnałów akustycznych w komputerowej analizie mowy. Wydawnictwo EXIT, Warszawa 2010.
10. Kamińska D.: Zastosowanie multimodalnej klasyfikacji w rozpoznawaniu stanów emocjonalnych na podstawie mowy spontanicznej. Warsztaty Doktoranckie, Lublin 2012.
11. Lewis M., Haviland-Jones J. M.: Psychologia emocji. Gdańskie Wydawnictwo Psychologiczne, Gdańsk 2005.
12. Ekman P., Davidson R.: Natura emocji – podstawowe zagadnienia. Gdańskie Wydawnictwo Psychologiczne, Gdańsk 1999.
13. Scherer K. R.: Vocal communication of emotion: A review of research paradigms. Speech Communication, No. 40, 2003, s. 227÷256.
14. Igras M., Wszołek W.: Pomiary parametrów akustycznych mowy emocjonalnej – krok ku modelowaniu wokalnej ekspresji emocji. Pomiary, Automatyka, Kontrola, Vol. 58, No. 4, 2012, s. 335÷338.
15. Igras M., Wszołek W.: Analiza percepcyjna i akustyczna mowy emocjonalnej aktorów. XIX Konferencja Inżynierii Akustycznej i Biomedycznej, Kraków-Zakopane 2012, s. 138.
16. Ziołko B., Miga B., Jadczyk T.: Semisupervised production of speech corpora using existing recordings. International 24. Seminar on Speech Production (ISSP'11), Montreal 2011.
17. Ziołko B., Ziołko M.: Przetwarzanie mowy. Wydawnictwo AGH, Kraków 2011.

Wpłynęło do Redakcji 9 stycznia 2013 r.

Abstract

We present a new database of Polish acted emotional speech patterns for purposes of analyzing emotional content in speech. Such databases are valuable and crucial resources for designing algorithms of automatic affects/attitude detection in speech, nevertheless they are difficult to collect.

Among most frequently used databases, the majority is based on emotions simulated by actors. The methodology allows to obtain good quality recordings of any desirable content and they do not involve ethical issues, though their main disadvantage is that the emotions are not genuine. Still, in comparison to other possible sources of emotional speech (tab. 1), they seem to be the best possible acoustic material to obtain. While there are many resources for foreign languages, there are only few for Polish speech, containing large enough database of recordings of sufficient signal quality.

Our database consists of good quality audio recordings of 12 speakers (6 male, 6 female), actors, drama students of volunteers. In total it consists of over 3.5 hours of recordings. For each speaker the same text content (tab. 2) – a set of words, dialogue utterances and a continuous text – was recorded, each time with one of the following emotions: joy, sadness, fear, surprise, anger, irony and neutral state as a reference. The paper describes the technical specification of the database. The process of data acquisition is described. The paper summarizes also the legal status and availability of the recordings for various use. The circumstances and variables affecting the recordings and critics of both advantages and disadvantages of the database type were discussed.

We present also the database applicability to many fields of speech processing, the range of possible research directions and applications for medical and commercial use.

Adresy

Magdalena IGRAS: AGH Akademia Górniczo-Hutnicza, Katedra Elektroniki,
al. Mickiewicza 30, 30-059 Kraków, Polska, migras@agh.edu.pl.
Bartosz ZIÓŁKO: AGH Akademia Górniczo-Hutnicza, Katedra Elektroniki,
al. Mickiewicza 30, 30-059 Kraków, Polska, bziolko@agh.edu.pl.