STUDIA INFORMATICA 2012
Volume 33 Number 2B (106)

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

AUDIOWIZUALNA BAZA NAGRAŃ MOWY POLSKIEJ

Streszczenie. Autorzy prezentują największą, audiowizualną bazę danych mowy polskiej i zarazem jedyną zrealizowaną w jakości HD. Artykuł przedstawia krótki opis podobnych baz dla innych języków oraz opis techniczny wykonanej bazy. Omówiono także napotkane wyzwania w trakcie realizacji bazy danych i jej planowane zastosowania.

Słowa kluczowe: rozpoznawanie mowy, przetwarzanie obrazów

AUDIOVISUAL DATABASE OF POLISH SPEECH RECORDINGS

Summary. The biggest audiovisual database of Polish speech (and the only one made in HD quality) is presented. The paper shortly introduces description of similar databases for other languages and the technical specification of the AGH database. The challenges met during the process of building the database are discussed along with the planned applications.

Keywords: speech recognition, image processing

1. Wprowadzenie

Rozpoznawanie mowy jest jednym z ważniejszych wyzwań współczesnej informatyki [1]. Wszystkie metody opierają się na danych statystycznych [2, 3, 4, 5]. Wykorzystanie analizy obrazu do rozpoznawania mowy wydaje się być kontrintuicyjne, jednakże poza Polską jest coraz częściej stosowane [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18]. Analiza ruchu warg jest tym istotniejsza, im głośniejszy jest szum w danym środowisku. W systemach opracowanych do działania w niskim SNR ma ona szczególne znaczenie. Kolejnym uzasadnieniem stosowania obrazu, jako dodatkowego źródła informacji, jest bimodalna natura procesu percepcji mowy u człowieka, jednocześnie przez zmysł słuchu i zmysł wzroku. Summerfield

[24] wymienia trzy powody, dla których obraz wspomaga rozpoznawanie mowy przez człowieka: zapewnia informacje o źródle dźwięku, dodatkowe informacje o położeniu artykulatorów i segmentach mowy. Podobnie więc efektywność działania systemów automatycznego rozpoznawania mowy może wzrosnąć dzięki uwzględnieniu informacji płynących z analizy wizualnej mówcy. Istotnie, w ostatnich latach widać dążenia do tworzenia inteligentnych interfejsów bazujących na fuzji komplementarnych informacji z wielu źródeł [10], zwłaszcza w środowiskach o wysokim poziomie szumu otoczenia [26].

Odpowiednia baza nagrań jest zasobem kluczowym dla badań nad audiowizualnym rozpoznawaniem mowy. Liczba zasobów dla języka angielskiego jest dość znaczna i wciąż wzrasta [19, 20, 21]. Zasoby różnią się rozmiarem, liczbą mówców, zawartością słowną oraz jakością nagrań. Jako przykłady można podać bazy XM2VTSDB [20], CUAVE [19], AVOZES [21], The GRID audiovisual sentence corpus [25]. Nie ma natomiast publicznie oferowanych korpusów audiowizualnych mowy polskiej. Wykonano jedynie nagrania w niskiej rozdzielczości i z niewielką zawartością słowną w ramach pracy doktorskiej M. Kubanka [23]. Ponadto, dotychczas analiza audiowizualna nie była szeroko stosowana w rozpoznawaniu mowy polskiej, a jedynie w analizie gestykulacji [22].

Korpus audiowizualny AGH stał się pierwszym tego typu zbiorem wzorców mowy języka polskiego, zawierającym blisko tysiąc słów, nagranym w jakości HD, otwierając nowe kierunki możliwych badań dla analizy obrazów twarzy i technologii mowy.

2. Audiowizualna baza nagrań mowy AGH

Korpus audiowizualny AGH zawiera nagrania audiowizualne mowy polskiej. Składa się z dobrej jakości nagrań twarzy (idealnie od frontu) 20 różnych mówiących osób (kobiet i mężczyzn) i transkrypcji wypowiedzi. Zawartość semantyczna nagrań każdego mówcy jest taka sama. Łączny czas trwania nagrań wynosi 200 minut (dodatkowo dołączone są nagrania testowe, niejednolitej jakości, 4 innych mówców – około 40 minut).

2.1. Parametry techniczne

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

Obraz rejestrowano za pomocą kamery JVC Everio GZ-HD500. Uzyskane nagrania mają postać plików .mts/avchd w standardzie H.264/MPEG-4 AVC, o parametrach: rozdzielczość HD 1920x1080, prędkość bitowa strumienia > 14 Mbps, liczba klatek 25/50 fps.

W celach prezentacyjnych zaleca się następujące parametry odtwarzania: monitor najlepiej z obsługą rozdzielczości 1920x1080, pamięć RAM minimum 2 GB, procesor minimum 3 GHz, karta graficzna ATI: modele z serii HD; NVIDIA: modele z technologią PureVideo HD, kodeki: K Lite Codec Pack: v. 7.9.2 (32 bit) / 5.4.0 (64 bit).

2.2. Rozmiar bazy danych audiowizualnych

Surowe dane po wykonaniu nagrań miały rozmiar (dla całego nagrania jednego mówcy): plik audio (w formacie .wav) – ok. 100 MB, plik wideo (.mts) – ok. 1,4 GB. Po przetworzeniu, polegającym na ponownej synchronizacji obrazu i dźwięku do jednego pliku w formacie .wmv, średnia wielkość plików wyjściowych wszystkich nagrań jednego mówcy (w formacie .wmv) wyniosła ok. 1,5 GB.

Nagranie każdego mówcy zostało podzielone na części tematyczne (osobno: zdania, teksty ciągłe, cyfry, polecenia) znajdujące się w osobnych plikach o wielkości: od ok. 15 MB (długość: ok.10 s) do ok. 600 MB (długość: ok. 5 min). Łączna wielkość zarchiwizowanych danych (całego korpusu) to 35 GB.

2.3. Warunki nagrywania

Nagrań dokonano w przeważającej części w ciągu dnia, przy naturalnym oświetleniu słonecznym. W nagraniach prowadzonych przy sztucznym oświetleniu zastosowano oświetlenie mieszane: oświetlenie ogólnie dostępne w salach zajęciowych, w których prowadzono nagranie, oraz, pomocniczo, oświetlenie miejscowe (światło rozproszone) dla zapewnienia zwiększenia doświetlenia rejonów twarzy niosących informacje.

Podczas nagrań mówca usytuowany był w pozycji siedzącej, na jasnym, jednolitym tle. Pomieszczenia (sale zajęciowe bez preferencyjnych warunków akustycznych), w których dokonano nagrań, miały poziom szumu ok. -60 dB.

2.4. Mówcy i treść

W nagraniach wzięło udział 24 mówców (11 kobiet, 13 mężczyzn) w wieku 20 – 26 lat. Po uprzednim zapoznaniu się z tekstami, zostali poproszeni o przeczytanie wyświetlanych kolejno tekstów (na ekranie komputera, na linii wzroku) w wyraźny (przy naturalnej dla siebie intonacji i tempie) sposób.

Dla każdego mówcy zarejestrowano nagrania o tej samej treści (około 10 minut dla każdego mówcy). Treść nagrań stanowią pojedyncze słowa (cyfry, polecenia sterujące), zdania dialogowe (160 zdań z naturalnych codziennych rozmów) oraz teksty ciągłe (7 tekstów: artykuły, definicje, fragmenty opowieści). Treść zaprojektowano pod kątem użycia w interfejsach człowiek-komputer oraz różnorodności typów wypowiedzi.

Rys. 1. Przykładowa ramka z nagrań
Fig. 1. Example of a frame from recordings

Rys. 2. Przykładowa ramka z nagrań
Fig. 2. Example of a frame from recordings

Rys. 3. Materiał informacyjny korpusu z ramkami nagrań każdego z mówców
Fig. 3. Graphical description of the corpus including frames of each speaker recordings

2.5. Metadane

W warstwie metadanych każde nagranie jest opisane za pomocą akronimu mówcy, oznaczenia jego płci oraz zawartości nagranej treści (osobno każdy tekst ciągły, zdania dialogowe, cyfry itd.). W dalszym etapie pracy z bazą nagrań metadane zostaną uzupełnione o anotacje czasowe wypowiadanych słów.

3. Doświadczenia z gromadzenia i archiwizacji danych

Zróżnicowane warunki oświetlenia oraz warunki akustyczne (hałas otaczającego środowiska, szum sprzętu, pogłos w pomieszczeniu) stanowią zarówno słabą, jak i mocną stronę korpusu, gdyż w rzeczywistym funkcjonowaniu systemów rozpoznawania mowy warunki idealne byłyby trudne do osiągnięcia. Jakość otrzymanego sygnału audio-wideo jest bardzo wysoka, co decyduje o dużej przydatności do badań nad algorytmami przetwarzania nagrań audiowizualnych, natomiast zastosowanie w interfejsach głosowych o szerokiej dostępności nie może zakładać powszechnego dysponowania tak dobrej jakości sygnałem, zwłaszcza w transmisji internetowej. Utrudnieniem, związanym z niedysponowaniem idealnym do tego zadania

sprzętem, była rejestracja osobno sygnału audio i sygnału wideo oraz wynikająca z tego konieczność synchronizacji podczas archiwizowania danych.

Użycie mowy czytanej zamiast spontanicznej umożliwiło otrzymanie nagrań o jednolitej treści dla każdego mówcy, ale wprowadziło też pewną dozę nienaturalności, która była niwelowana przez uprzednie zapoznanie się mówców z treścią tekstów. Ostateczny wynik w zakresie naturalności mowy, oszacowany subiektywną oceną percepcyjną, jest różny dla różnych osób. Na różnorodność nagrań miały też wpływ zarówno indywidualne różnice w intonowaniu i wyraźności mówienia, jak i aktualny stan emocjonalny mówców.

Rozwijając istniejący lub projektując nowe korpusy audiowizualne, należy pamiętać o dostosowaniu do danego zastosowania zarówno treści słownej, jak i jakości nagrań. Nie bez znaczenia jest również przygotowanie mówców i sposób wymawiania przez nich tekstów (naturalny a wyraźne intonowanie; mowa spontaniczna a teksty czytane).

4. Aspekty prawne

Nagrania posiadają uregulowaną sytuację prawną, umożliwiającą ich:

- **Przetwarzanie:** Zgoda mówców na wykorzystanie naukowe oraz przetwarzanie nagrań w systemach informatycznych technologii mowy, w tym komercyjnych.
- **Prezentowanie publiczne:** Zgoda na anonimowe (bez podawania imienia i nazwiska nagranej osoby) odtwarzanie na konferencjach, wykładach i prezentacjach systemów technologii mowy.

5. Podsumowanie

W artykule opisano nowopowstałą bazę wzorców audiowizualnych języka polskiego. Ma ona istotne znaczenie w dalszym rozwoju badań nad wielomodalnymi technikami rozpoznawania mowy oraz potencjalnie nad analizą obrazów twarzy i produkcji graficznych wirtualnych doradców z syntezą mowy, w których ruch ust będzie wiarygodnie adaptował się do syntezowanej mowy. Opracowana baza danych umożliwia analizę odpowiadających sobie wzorców sygnałów akustycznego i wizualnego, występujących podczas aktu mowy. Jej stosowanie może prowadzić do rozwoju technik łączenia danych ze strumieni audio i wideo w systemach rozpoznawania mowy. Baza może być także wykorzystana w produkcji systemów „czytania z ruchu ust". Nagrania umożliwią badania podstawowe w zakresie analizy obszarów twarzy przenoszących informacje (usta, szczęki, policzki) i artefaktów zakłócających rozpoznanie mowy oraz analizę zmian mimiki twarzy towarzyszących mówieniu.

6. Opis planowanych prac badawczych

Na dalszym etapie badań, w celu przystosowania bazy nagrań do wykorzystania w audiowizualnym przetwarzaniu mowy, zostanie ona uzupełniona o transkrypcje fonetyczne oraz anotacje czasowe każdego nagrania oraz anotacje obszarów ROI. W celu usprawnienia procesu anotowania korpusu zostaną opracowane algorytmy półautomatycznej, automatycznej anotacji nagrań. Tak opracowany korpus może zostać wykorzystany do opracowania i testowania algorytmów analizy twarzy i rozpoznawania mowy pod kątem wielomodalnego przetwarzania mowy polskiej oraz zaawansowanych interfejsów człowiek-komputer.

Kolejne fazy procesu analizy będą obejmowały:

- detekcję twarzy i obszarów przenoszących informację (ROI): ust, szczęk, policzków,
- algorytmy ekstrakcji cech oraz metody określające obszary zawierające informację przydatną do przetwarzania mowy, a także sposoby parametryzacji wybranych regionów obrazu,
- różne techniki klasyfikacji danych wielostrumieniowych i algorytmy redukcji wymiarowości wektorów cech – w celu poprawy jakości klasyfikacji danych,
- strategię łączenia danych pochodzących ze strumieni audio i wideo.

W przyszłości zostanie zweryfikowana możliwość zastosowania dla mowy polskiej algorytmów audiowizualnego przetwarzania mowy stosowanych dla innych języków. Na bazie uzyskanej wiedzy na temat możliwości i skuteczności wykorzystania poszczególnych algorytmów przetwarzania danych audiowizualnych do analizy mowy polskiej, zostaną wybrane, zoptymalizowane i zaimplementowane najbardziej efektywne algorytmy, w celu stworzenia systemu audiowizualnego rozpoznawania mowy oraz wykorzystania w dalszych badaniach nad tworzeniem zaawansowanych interfejsów komunikacji człowiek-komputer.

Praca naukowa finansowana ze środków na naukę jako projekt badawczy NCBiR 0021/R/D2/201/01 O ROB 0021 01/ID 21/2 i z działalności statutowej.

BIBLIOGRAFIA

1. Ziółko B., Ziółko M.: Przetwarzanie mowy, Wydawnictwa AGH, Kraków 2011.
2. Demenko G., Grocholewski S., Klessa K., Ogórkiewicz J., Wagner A., Lange M., Śledziński D., Cylwik N.: JURISDIC – Polish speech database for taking dictation of legal texts. Materiały konferencyjne of the International Conference on Language Resources and Evaluation, 2008, s. 1280÷1287.

3. Young S., Evermann G., Gales M., Hain T., Kershaw D., Moore G., Odell J., Ollason D., Povey D., Valtchev V., Woodland P.: HTK Book. UK: Cambridge University Engineering Department, 2005.
4. Lamere P., Kwok P., Gouvea E., Raj B., Singh R., Walker W., Wolf, P.: The CMU Sphinx-4 speech recognition system. Sun Microsystems, 2004.
5. Ziółko M., Gałka J., Ziółko B., Jadczyk T., Skurzok D., Mąsior M.: Automatic Speech Recognition System Dedicated for Polish. Show and tell session, materiały konferencyjne Interspeech, Florencja 2011.
6. Terry L. H., Katsaggelos A. K.: A phone-viseme dynamic Bayesian network for audio-visual automatic speech recognition. Materiały konferencyjne ICPR, 2008.
7. Adjoudani A., Benoit C.: On the integration of auditory and visual parameters in an HMM-based ASR, [in:] Stork D. G., Hennecke M. E. (eds.): Speech reading by Humans and Machines: Systems and Applications. Springer-Verlag, Berlin, Germany 1996, s. 461÷472.
8. Basu S., Oliver N., Pentland A.: 3D modeling and tracking of human lip motions. Materiały konferencyjne International Conference on Computer Vision, Mumbai, India 1998, s. 337÷343.
9. Borgstrom B. J., Alwan A.: A Low-Complexity Parabolic Lip Contour Model With Speaker Normalization for High-Level Feature Extraction in Noise-Robust Audiovisual Speech Recognition. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, Vol. 38, No. 6, 2008, s. 1273÷1280.
10. Chen T.: Audiovisual speech processing. IEEE Signal Processing Magazine, Vol. 18, No. 1, 2001, s. 9÷21.
11. Dupont S., Luettin J.: Audio-visual speech modeling for continuous speech recognition. IEEE Transactions on Multimedia, Vol. 2, No. 3, 2000, s. 141÷151.
12. Gagnon L., Foucher S., Laliberte F., Boulianne G.: A simplified audiovisual fusion model with application to large-vocabulary recognition of French Canadian speech. Canadian Journal of Electrical and Computer Engineering, Vol. 33, No. 2, 2008, s. 109÷119.
13. Gowdy J., Subramanya A., Bartels C., Bilmes J.: DBN based multi-stream models for audio-visual speech recognition. Materiały konferencyjne IEEE International Conference on Acoustic, Speech and Signal Processing, 2004, s. 993÷996.
14. Gurban M., Thiran J.: Audio-visual speech recognition with a hybrid SVM-HMM system. Materiały konferencyjne 13th European Signal Processing Conference EUSIPCO, Antalya, Turkey 2005.
15. Heckmann M., Berthommier F., Kroschel K.: A hybrid ANN/HMM audiovisual speech recognition system. Materiały konferencyjne International Conference on Auditory-Visual Speech Processing, Aalborg, Denmark 2001, s. 190÷195.

16. Huang F. J., Chen T.: Consideration of Lombard effect for speech reading. Materiały konferencyjne Workshop on Multimedia Signal Processing, Cannes 2001, s. 613÷618.
17. Huang J., Potamianos G., Neti C.: Improving audio-visual speech recognition with an infrared headset. Materiały konferencyjne Workshop on Audio-Visual Speech Processing, St. Jorioz, France, 2003, s. 175÷178.
18. Zdansky J., Chaloupka J., Nouza J.: Joint audio-visual processing, representation and indexing of TV news programmes. 2008 IEEE 10th Workshop on Multimedia Signal Processing, 2008, s. 960÷965.
19. Patterson E. K., Gurbuz S., Tufekci Z., Gowdy J. N.: CUAVE: A new audiovisual database for multimodal human-computer interface research. Materiały konferencyjne International Conference on Acoustics, Speech and Signal Processing, Orlando 2002, s. 2017÷2020.
20. Messer K., Matas J., Kittler J., Jonsson K.: XM2VTSDB: The Extended M2VTS Database. Materiały konferencyjne Second International Conference on Audio and Video-based Biometric Person Authentication, 1999, s.72÷77.
21. Goecke R., Millar J. B.: The Audio-Video Australian English Speech Data Corpus AVOZES, 2004.
22. Karpiński M., Jarmołowicz-Nowikowa E., Malisz Z., Szczyszek M., Juszczyk K.: Rejestracja, transkrypcja i tagowanie mowy oraz gestów w narracji dzieci i dorosłych. Investigationes Linguisticae, Vol. XVI; Poznań 2008.
23. Kubanek M.: Metoda rozpoznawania audio-wideo mowy polskiej w oparciu o ukryte modele Markowa. Praca doktorska Politechniki Częstochowskiej, 2005.
24. Summerfield A. Q.: Some preliminaries to a comprehensive account of audio-visual speech perception, [in:] Dodd B., Campbell R. (eds.): Hearing by Eye: The Psychology of Lip-Reading. Lawrence Erlbaum Associates, London, United Kingdom 1987, s. 3÷51.
25. Barker J., Cooke M.: Modelling speaker intelligibility in noise. 2007.
26. Shivappa S. T., Trivedi M. M., Rao B. D.: Audiovisual Information Fusion in Human--Computer Interfaces and Intelligent Environments: A Survey. Journal: Proceedings of The IEEE-PIEEE, 2010.

Wpłynęło do Redakcji 5 stycznia 2012 r.

Abstract

The majority of existing automatic speech recognition systems are based on audio stream, while fusion of both audio and visual data could highly increase their efficiency especially in low-SNR environments. However, there has been a lack of such resources for Polish speech, containing large enough database of recordings of sufficient signal quality.

For purposes of audiovisual speech processing, the biggest audiovisual database of Polish speech was created. As the only one, it is made in HD quality. In total it consists of almost 4 hours of recording of read speech of 24 speakers (for each the same text content). Examples of frames from recordings are attached (fig.1 – 2). The paper introduces the technical specification and other parameters (speakers, semantic content, files size) of the AGH database. The process of data acquisition was described, as well as acoustic and lighting conditions. The paper summarizes also the legal status and availability of the recordings for various use.

The challenges met during the process of building the database are discussed. Analysis of the circumstances affecting the recordings quality and critics of both advantages and disadvantages of the database diversity led to conclusions that can be helpful in further development of the database and designing new audiovisual speech corpora.

The AGH database is applicable to many fields of audiovisual data processing. The range of possible research directions and planned applications is presented.

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.
Tomasz JADCZYK: AGH Akademia Górniczo-Hutnicza, Katedra Elektroniki,
al. Mickiewicza 30, 30-059 Kraków, Polska, jadczyk@agh.edu.pl.