

Baza nagrań głosowych z dostępem poprzez Internet

The system architecture of a speech database via the WWW

Andrzej Izworski, Piotr Pracuch, Jarosław Bułka, and Ireneusz Wochlik

Katedra Automatyki, Akademia Górniczo-Hutnicza w Krakowie
izwa@agh.edu.pl

STRESZCZENIE

W pracy przedstawiono założenia i sposób implementacji realizowanej bazy danych zawierającej nagrania dźwiękowe. Projektując bazę założono, że źródłem nagrań będą rozmowy prowadzone przez telefon komórkowy. Zarejestrowane rozmowy mogą być następnie przesyłane do centrum zapisowego za pomocą aplikacji zainstalowanej w połączeniu z przeglądarką internetową. Transfer danych do przeglądarki może następować za pomocą transmisji radiowej lub za pomocą połączenia kablowego. Przechowywane nagrania są następnie edytowane i anotowane przez dwie grupy użytkowników: techników i ekspertów. Każda z grup posiada inne prawa dostępu. Dodatkowo bazę danych uzupełniono o specjalizowane oprogramowanie pozwalające na wizualizację i odsłuchiwanie oraz indeksujące i udostępniające nagrania lub ich wyselekcjonowane fragmenty. Wbudowano możliwość korzystania z procedur rozpoznawania elementów mowy, rozpoznawania w nagraniach uprzednio zdefiniowanych słów kluczowych lub fraz, rozpoznawania rozmówców. Ze względu na dostęp do bazy danych poprzez Internet wdrożono odpowiednie środki bezpieczeństwa przekazywanych informacji. Przyjęte rozwiązania ułatwiają wdrażanie systemu w dużych, rozproszonych instytucjach, nie nakładając jednocześnie ograniczeń na rozmiary bazy danych.

ABSTRACT

The aim of the paper is to introduce basic assumptions regarding design and architecture of the sound recordings database. The main assumption concerning database design expects data to come from mobile phone conducted conversations. Recorded conversations can be sent to the web-based central application through a web browser. Data from mobile phones, to computers that can access central application through a web browser, can be transferred by radio transmission or cable connection. Recordings stored in central database can be edited and annotated by two groups of users: technicians and experts – each group having different access rights. The central database has been extended by addition of specialized visualization and indexing software for sound recordings or their selected parts. Software includes sound recognition procedures with voice recognition, able to identify keywords, phrases or speakers. The database, being established in public network, required security instruments to be implemented, therefore to restrict access and secure stored information. The system architecture simplifies implementation in large, distributed environment, while not imposing any restrictions on the size of the database.

1 Wstęp

Celem pracy jest prezentacja aktualnie realizowanego systemu do akwizycji, wizualizacji, przetwarzania, automatycznego rozpoznawania oraz opisu nagrań dźwiękowych

umożliwiającego pracę z danymi w środowisku zdecentralizowanym, gdzie użytkownicy znajdują się w różnych lokalizacjach. Realizacja dostępu do danych zapewnia elastyczność w dostępie do zasobów systemu niezależnie od lokalizacji samego użytkownika.

Przyjęto, iż źródłem nagrań będą rozmowy prowadzone przez telefon komórkowy i następnie zarejestrowane przy pomocy telefonu. Możliwe jest również korzystanie z innych kanałów akwizycji nagrań. Zarejestrowane rozmowy mogą być następnie przesyłane do centrum zapisowego za pomocą aplikacji zainstalowanej w połączeniu z przeglądarką internetową. Transfer danych do przeglądarki może następować w sposób całkowicie automatyczny za pomocą transmisji radiowej lub za pomocą połączenia kablowego. Użycie Internetu jako medium komunikacji zapewnia dostęp do dobrze ustandaryzowanych protokołów, co z kolei ułatwia dostęp do zasobów systemu. Przeglądarka internetowa, jako główne narzędzie pracy z systemem jest zainstalowana na prawie każdym komputerze osobistym, co ułatwia wdrażanie systemu w dużych, rozproszonych instytucjach.

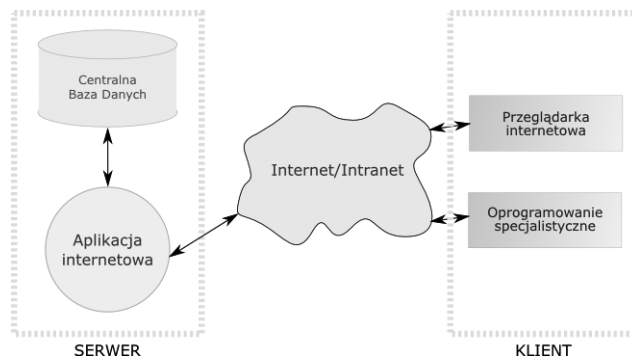
Przechowywane nagrania są następnie edytowane i anotowane przez dwie grupy użytkowników: techników i ekspertów. Każda z grup posiada inne prawa dostępu. Wdrożenie systemu umożliwia zdalne i racjonalne wykorzystanie cennych zasobów ludzkich w postaci specjalistów z zakresu analizy nagrań. Z uwagi na systematyczne przysyłanie nagrań w obrębie zainteresowanych użytkowników konieczne stało się wdrożenie odpowiednich środków bezpieczeństwa przekazywanych informacji. Przyjęte rozwiązania ułatwiają wdrażanie systemu w dużych, rozproszonych instytucjach, nie nakładając jednocześnie ograniczeń na rozmiary bazy danych i pozwalając na zdalne wykorzystanie ekspertów.

Dodatkowo bazę danych uzupełniono o specjalizowane oprogramowanie pozwalające na wizualizację i odsłuchiwanie oraz indeksujące i udostępniające nagrania lub ich wyselekcjonowane fragmenty. Wbudowano możliwość korzystania z procedur rozpoznawania elementów mowy, rozpoznawania w nagraniach uprzednio zdefiniowanych słów kluczowych lub fraz, rozpoznawania rozmówców. Wyniki rozpoznawania mogą być dodane do anotacji tworząc pełny opis przechowywanych w bazie wypowiedzi.

2. Architektura systemu

Potencjalny użytkownik złożonego, rozproszonego systemu informatycznego to najczęściej osoba nie posiadająca gruntownego informatycznego, będąca jednak ekspertem w swojej dziedzinie i traktująca oprogramowanie jako narzędzie codziennej pracy. Uświadomienie tego faktu zmusza twórców aplikacji informatycznych do szczególnej dbałości o stosowane rozwiązania z punktu widzenia przyjaznej komunikacji człowiek-komputer, dużej łatwości definiowania zadań stawianych przed aplikacją oraz łatwość wprowadzania modyfikacji bez udziału specjalistów informatyki ad hoc przez użytkowników wytworzone i wdrożone aplikacje. Powyższe założenia wpłynęły bardzo silnie na postać prezentowanego systemu do przechowywania i przetwarzania nagrań dźwiękowych.

System zrealizowany został w architekturze klient-serwer, gdzie klienci łączą się z serwerem w celu uzyskania dostępu do zasobów przechowywanych w centralnej bazie danych z użyciem otwartych, ustandaryzowanych protokołów (ryc. 1). Kolejnym istot-



Rycina 1. Architektura systemu.

nym elementem w przypadku wykorzystania w pracy systemów przeglądarki internetowej, jest zgodność oprogramowania ze standardami W3C, które ujednolicają sposób prezentowania informacji. Tworzenie systemów informatycznych w oparciu o otwarte standardy gwarantuje ich elastyczność, łatwość komunikacji z innymi systemami oraz gotowość do dalszej rozbudowy. Ten ostatni warunek jest jednym z kluczowych dla projektu ze względu na bardzo szybki rozwój technologii i coraz to nowszych badań dających wyniki w postaci cyfrowej, które powinny być docelowo obsługiwane przez projektowany system. System został zaprojektowany w architekturze wielowarstwowej, a dobór komponentów poszczególnych warstw umożliwia ich klastrowanie.

Zaprojektowany w systemie trójwarstwowym system przewiduje użycie dwóch rodzajów klientów:

- „cienki klient” w postaci przeglądarki internetowej używany do przeglądania i przeszukiwania zasobów zgromadzonych w centralnej bazie danych za pośrednictwem aplikacji internetowej oraz akwizycji danych,
- „gruby klient” w postaci oprogramowania specjalistycznego, które łącząc się z aplikacją internetową za pomocą mechanizmów webservice otrzymuje dostęp do nagrań dźwięku w celu jego analizy i obróbki.

3. Akwizycja danych

Przyjęto następujący trójstopniowy schemat akwizycji i przetwarzania danych:

- źródło dźwięku – telefon komórkowy,
- wprowadzanie nagrania i informacji do systemu – panel webowy bądź aplikacja zaimplementowana w telefonie komórkowym,
- przetwarzanie, wizualizacja, analiza nagrań – poprzez centralną bazę danych.

Źródłem nagrań dźwięku może być telefon komórkowy, dyktafon lub jakiegokolwiek inne urządzenie posiadające możliwość przesłania zapisu dźwięku na komputer posiadający dostęp do systemu. Nagranie jest następnie transferowane do komputera pełniącego rolę klienta. Możliwy jest bezprzewodowy sposób transferu (automatyczna syn-

chronizacja aplikacji w telefonie z aplikacją zainstalowaną na komputerze klienckim) lub też zestawienie połączenia za pomocą kabla. Pliki zawierające zapis dźwięku są w dalszej kolejności uzupełniane o etykietę (meta dane) zawierającą podstawowe informacje o nagraniu. Wprowadzane metadane pozwalają między innymi na identyfikację zapisu, dzięki nim możliwe jest również wyszukiwanie zapisów według wybranych kryteriów. Akwizycją nagrania i jego wstępnym (podstawowym) zapisem zajmuje się specjalizowana aplikacja zainstalowana w komputerze klienckim. Następnie za pomocą przeglądarki internetowej pliki z nagraniem i jego opisem są przesyłane do centralnej bazy danych.

4. Analiza zapisu dźwięku

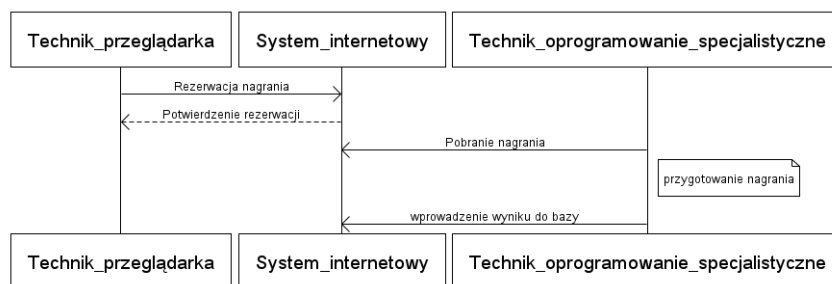
Zgromadzone w systemie dane mogą być odsłuchiwane, wizualizowane i sortowane oraz raportowane przez każdego licencjonowanego użytkownika. Często jednak występuje konieczność przeprowadzenia zaawansowanej analizy nagrań przez eksperta. Implementowany system wprowadza taką możliwość opisu i analizy zapisu nagrań, realizowaną przez oprogramowanie klienckie po stronie użytkownika. Umiejscowienie oprogramowania analizującego na maszynach klienckich aplikacji internetowej pozwala zmniejszyć obciążenie części serwerowej.

Oprogramowanie specjalistyczne do pracy nad zapisem dźwięku stanowi grubego klienta do systemu internetowego – komunikuje się bezpośrednio z aplikacją internetową z pominięciem przeglądarki pobierając dane źródłowe i wysyłając dane wynikowe.

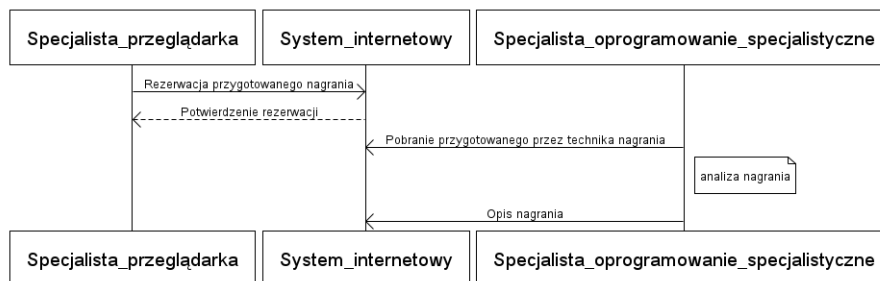
Założono potrzebę istnienia hierarchicznej struktury użytkowników wykonujących analizę zapisu dźwięku, a mianowicie ich podział na dwie kategorie – techników i ekspertów:

- technicy – są to użytkownicy przygotowujący zapis dźwięku do analizy, usuwający fragmenty nieistotne z punktu widzenia specjalisty, decydujący o merytorycznym znaczeniu nagrania, itp. (ryc. 2),
- eksperci – dokonują opisu zapisu dźwięku przygotowanego przez technika (ryc. 3).

Poprzez rozdzielenie zadań technika i eksperta optymalizowane jest wykorzystanie specjalistów, którzy odciążeni są z przygotowania zapisu i skupiają się wyłącznie na je-



Rycina 2. Schemat analizy nagrania dźwiękowego przez technika – zewnętrznego specjalistę.



Rycina 3. Schemat analizy nagrania dźwiękowego przez eksperta – zewnętrznego specjalistę.

go analizie. W system wbudowane zostały mechanizmy dbające o to, aby technicy i eksperci nie podejmowali pracy nad tym samym zapisem dzięki mechanizmowi rezerwacji zadań. Przed przystąpieniem do pracy, technicy i eksperci za pomocą panelu aplikacji internetowej dokonują rezerwacji zasobu, jakim jest zapis dźwięku, dzięki czemu mogą rozpocząć jego obróbkę w oprogramowaniu specjalistycznym, uniemożliwiając jednocześnie pracę nad tym zapisem innym użytkownikom. Taka strategia eliminuje sytuacje, gdy dwóch lub więcej użytkowników wykonuje identyczną pracę – należy pamiętać, że w dużych systemach, gdzie użytkownicy są rozproszeni geograficznie, nie zawsze istnieje możliwość bezpośredniej koordynacji ich działań poza systemem.

Wynik działania oprogramowania specjalistycznego przesyłany jest do systemu i dołączany do metadanych. Wynik staje się dostępny dla użytkowników poprzez przeglądanie i wyszukiwanie w danych zgromadzonych w centralnej bazie danych.

5. Podsumowanie

Proponowane rozwiązania projektowe pozwoliły na stworzenie elastycznego systemu zbudowanego w oparciu o architekturę wielowarstwową co pozwala na dużą łatwość wprowadzania nowych funkcjonalności. Architektura systemu pozwala jednocześnie na elastyczne konfigurowanie i łatwe modyfikowanie rozwiązań informatycznych definiowanych przez użytkownika nie posiadającego przygotowania informatycznego.

Istotnym novum proponowanego systemu jest wyraźne rozgraniczenie ról i możliwości użytkowników oraz ich podział na tych którzy wprowadzają nowe wartości do opisów nagrań (anotacja) oraz na tych którzy jedynie korzystają w swojej pracy z uprzednio przetworzonych nagrań.

Kolejną ważną pozytywną cechą proponowanego rozwiązania jest wmontowanie w aplikację mechanizmów samouczenia się aplikacji i automatycznego zbierania danych. Zebrana w trakcie eksploatacji baza danych pozwoli na lepsze wykorzystanie systemu bez ponoszenia dużych nakładów pracy i środków na wprowadzanie danych.

Proponowane rozwiązania szczegółowe mają również walory naukowe i mogą stanowić istotny przyczynek do postępu w badaniach nad automatycznym rozpoznawaniem mowy.

BIBLIOGRAFIA

- [1] Bombien L., Cassidy S., Harrington J., John T., Palethorpe S. (2006) *Recent Developments in the Emu Speech Database System*, Proceedings of the Australian Speech Science and Technology Conference, Auckland, December 2006.
- [2] Chmurzyńska K., Radkowski P., Izworski A., Orzechowski T., (2005) *Mobile and Internet-based system for computer aided diagnosis with auditory brainstem responses*, Proceedings of the International Conference on Intelligent Systems, Kuala Lumpur, 1–3 December 2005.
- [3] Czyżewski A., (1998) *Dźwięk cyfrowy*, AOW EXIT, Warszawa.
- [4] Demenko G., Grocholewski S., Klessa K., Ogórkiewicz J., Lange M., Śledziński D., Cylwik N., (2008) *Jurisdic-Polish Speech Database for taking dictation of legal texts*. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 28–30 May 2008 Marrakech, Morocco.
- [5] Izworski A., Kochanek K., Bulka J., Sliwa L., Wochlik I., (2004) *Internet-wise acquisition and multimedia data visualisation for ABR*, Electronic Journal of Pathology and Histology, 2nd International Conference on Telemedicine and multimedia communication, October 8–9, 2004 Warsaw-Kajetany, Poland.
- [6] Siemund R., Hoeghe H., Kunzmann S., Marasek K. (2000) *SPEECON – speech data for consumer devices*, In Proceedings of LREC.
- [7] Tadeusiewicz R. (1998) *Sygnal mowy*, WKiŁ, Warszawa.