Streszczenie

Przedmiotem pracy są metody **przemiany głosu** – technologii mowy, która polega na transformacji cech osobniczych zawarych w sygnale mowy na cechy wybranego innego mówcy, przy zachowaniu nienaruszonej struktury wypowiedzi. Celem takiego przetwarzania jest wywarcie wrażenia, że wypowiedź pochodzi od innego mówcy niż w rzeczywistości. W pracy skupiono się na cechach segmentalnych (krótkoczasowych) głosu, które nadają się do transformacji w czasie rzeczywistym. Należą do nich wysokość tonu i barwa głosu.

Aby możliwe było przetransformowanie barwy głosu, którą w niniejszej pracy utożsamia się z obwiednią widmową sygnału, potrzebna jest ilościowa reprezentacja tejże wielkości przy pomocy wektora współczynników i procedura ich estymacji. W tym celu stosuje się różne modele obwiedni widmowej sygnału mowy. Badaniom poddano trzy, najczęściej wykorzystywane w technologiach mowy: liniowy model predykcyjny (reprezentowany na dwa sposoby: logarytmicznymi stosunkami przekrojów oraz częstotliwościami widma prążkowego), model homomorficzny i tzw. mel-cepstralny model przedziałami liniowy.

Transformacja barwy głosu polega na odwzorowaniu współczynników obwiedni widmowej z przestrzeni fonetycznej mówcy źródłowego w przestrzeń mówcy docelowego, czyli jest problemem nieliniowej regresji w wielu wymiarach. Do realizacji tego zadania wykorzystano dwie alternatywne metody uczenia maszynowego: sztuczną sieć neuronową (SSN) o topologii wielowarstwowego perceptronu oraz metodę wektorów podpierających (SVM). W procesie uczenia wykorzystano równoległe zbiory uczące, utworzone na podstawie korpusu nagrań mowy polskiej *CORPORA*. Zbiory te otrzymano w procedurze dopasowania wypowiedzi w czasie, której celem jest powiązanie odpowiadających sobie fonetycznie segmentów w analogicznych wypowiedziach pochodzących od różnych mówców. Dzięki temu możliwe jest znalezienie statystycznych korelacji między cechami obwiedni widmowej mówcy źródłowego i docelowego, a w efekcie znalezienie odwzorowania, które zmieni cechy osobnicze, zachowując strukturę fonetyczną wypowiedzi.

Wysokość tonu jako wielkość skalarna okazała się łatwiejsza do transformacji. Do realizacji tego zadania zaproponowano dwie metody: parametryczną i nieparametryczną.

Metoda parametryczna polega na jawnym wyznaczeniu wysokości tonu jako jednej z cech opisujących wzbudzenie i jej liniowej bądź afinicznej transformacji, a następnie wygenerowaniu syntetycznego wzbudzenia o zmienionej wysokości tonu. W tym celu utworzono parametryczny system analizy-syntezy wzbudzenia, inspirowany wokoderem MELP. Wprowadzono w nim oryginalną metodę estymacji dźwięczności jako parametru ciągłego (a nie

dyskretnego, jak w rozwiązaniach konwencjonalnych), co pozwoliło uniknąć konieczności segmentacji lub klasyfikacji segmentów ze względu na dźwięczność, a przez to na realizację przemiany w sposób ciągły i jednorodny.

Metoda nieparametryczna wykorzystuje do transformacji wysokości tonu wokoder fazowy i nie wymaga jawnego wyznaczenia wysokości, a jedynie ustalenia stosunku, w jakim ton ma być przeskalowany. Chociaż wokoder fazowy jest rozwiązaniem znanym już od długiego czasu, brak jak dotąd doniesień o jego zastosowaniu w przemianie głosu.

Uzyskane różnymi metodami wyniki przemiany głosu zostały poddane ocenie w formalnych eksperymentach odsłuchowych celem porównania ich skuteczności i jakości. W odsłuchach wzięli udział niezależni słuchacze ochotnicy, w większości studenci Interdyscyplinarnego Studium Doktoranckiego w IPI PAN. W doświadczeniach tych wykorzystano próbki dźwiękowe uzyskane przez transformację trzech głosów korpusu *CORPORA* (mężczyzny, kobiety i chłopca). Pod uwagę wzięto metodę parametryczną transformacji tonu oraz różne kombinacje wszystkich czterech wymienionych reprezentacji obwiedni widmowej oraz obu metod uczenia maszynowego (SSN i SVM). Wyniki dowodzą skuteczności badanych metod i wskazują na znaczną stratę jakości na etapie transformacji obwiedni widmowej. Na tle danych z literatury, skuteczność plasuje się w przedziale typowo osiąganym przez inne rozwiązania, natomiast jakość nieco niżej niż większość opisywanych systemów.

Wybrane metody i rozwiązania zostały zrealizowane w postaci programów komputerowych, pozwalających na transformację głosu w czasie rzeczywistym. Opóźnienia systemu są na tyle małe, że pod względem technicznym możliwe jest jego wykorzystanie praktyczne.

Abstract

The subject of the thesis are the methods of **voice conversion** - a speech technology which consists in the transformation of speaker identity features contained in the speech signal towards the features of a selected, target speaker, while keeping the structure of the utterance intact. The goal is thus to create perceptual impression that a given utterance is spoken by a speaker different than original one. The presented work focuses on modification of segmental (short-time) features of speech which are suitable for transformation in real time. This includes the pitch and the *timbre*, or the spectral colouring of the signal.

In order to transform the timbre, which in this work is associated with the frequency envelope of the signal spectrum, a quantitative representation of its shape using a vector of parameters and a method of their estimation are needed. To this end, various models of the spectral envelope can be employed. Three of them were included in this study: the linear predictive model (with two alternative representations, using either logarithmic area ratios or line spectral frequencies), the homomorphic („cepstral") model and the piecewise linear model based on mel-frequency cepstral coefficients.

Transformation of timbre is a mapping of spectral envelope coefficients from the phonetic space of source speaker into target speaker's space and can be viewed as the task of multidimensional nonlinear regression. Two methods of machine learning were used for this purpose: the artificial neural network (ANN) with multilayer perceptron topology and the support vector method (SVM). For training, parallel data sets were used, created from the *CORPORA* corpus of spoken Polish. The datasets were obtained in a time alignment procedure, the aim of which was to bind phonetically analogous segments in corresponding utterances of different speakers and thus level prosodic differences. As a result, statistical correlations between spectral envelope features of the source and target speaker can be found, and consequently used for finding the transformation function, which changes identity features, while preserving the phonetic structure of the utterance.

The pitch, being a scalar quantity, proved easier to transform. For this task, two methods have been proposed: a parametric and a nonparametric one.

The parametric method consists in explicitly determining the fundamental frequency among other excitation parameters, its transformation using linear or affine function, followed by resynthesis of synthetic excitation with modified pitch. To this end, a parametric speech analysis-synthesis system was created, which is inspired by the MELP vocoder. An original method is introduced for the estimation of the degree of voicing as a continuous quan-

tity (as opposed to discrete voiced-unvoiced states as used in conventional approaches). It allows to avoid the need for segmentation or classification of speech frames based on their voicing, and thus to process speech in a continuous and homogeneous manner.

The nonparametric method uses the phase vocoder to transform the pitch and does not require an explicit estimation of the fundamental period or frequency. Only the ratio in which the pitch is to be rescaled must be defined. While the phase vocoder has been known for a ling time, little had been published on its use in voice conversion so far.

The effects of processing of speech using various methods of voice conversion were assessed in formal listening tests in order to compare their quality and efficiency. Independent listeners (mostly doctoral students of the Interdisciplinary Doctoral Studies at IPI PAN) were invited into several listening sessions. Speech samples obtained by processing voices of three speakers from *CORPORA* (a man, a woman and a boy) were used in the experiments. A parametric approach to pitch transformation was applied with various combinations of the four studied representations of the spectral envelope and the two mentionend machine learning methods (ANN and SVM). The results prove successfull conversion with all considered methods while indicating a significant loss of quality, which is predominantly due to envelope transformation. Considering published results from literature as reference, the examined methods are comparable in terms of efficiency and below average in terms of quality.

Selected methods and solutions were implemented as computer programs which allow the transformation of voice in real time. The latencies are small enough to allow practical applications from the technical point of view.