VOICE CONVERSION USING K-HISTOGRAMS AND RESIDUAL AVERAGING

A.J. URIZ[†], P.D. AGÜERO[‡], J.C. TULLI[‡], J. CASTIÑEIRA MOREIRA[†], E.L. GONZÁLEZ[‡] and A. BONAFONTE[§]

† *CONICET -Facultad de Ingeniería, Universidad Nacional de Mar del Plata, 7600 Mar del Plata, Argentina. ajuriz@conicet.gov.ar*
‡ *Facultad de Ingeniería, Universidad Nacional de Mar del Plata, 7600 Mar del Plata, Argentina, pdaguero@fi.mdp.edu.ar*
§ *Universitat Polytècnica de Catalunya, Barcelona, Spain*.

Abstract — **The main goal of a voice conversion system is to modify the voice of a source speaker, in order to be perceived as if it had been uttered by another specific speaker. Many approaches found in the literature convert only the features related to the vocal tract of the speaker. Our proposal is to convert those characteristics, and to process the signal passing through the vocal chords. Thus, the goal of this work is to obtain better scores in the voice conversion results.**

Keywords— **Voice Conversion, K- Histograms, Residual Conversion, Voice Synthesis.**

I. INTRODUCTION

The primary goal of voice conversion systems is to modify the voice of a source speaker, in order to be perceived as if it had been uttered by another specific speaker, the target speaker. For this purpose, relevant features of the source speaker are identified and replaced by the corresponding features of the target speaker.

Several voice conversion techniques have been proposed since the problem was first formulated in 1988. In this year, Abe *et al.* (1988) proposed to convert voices by mapping codebooks created from a parallel training corpus. Since then, many authors tried to avoid spectral discontinuities caused by the hard partition of the acoustic space, by means of fuzzy classification or frequency axis warping functions. The introduction of statistical methods based on gaussian mixture models (GMM) for spectral envelope transformation was an important breakthrough in voice conversion (Kain, 2001; Stylianou *et al.*, 1998). In these approaches, the acoustic space of speakers is partitioned into overlapping classes and the weighted contribution of all classes is considered when transforming acoustic vectors. As a result, the spectral envelopes are successfully converted without discontinuities, but as a downside, the quality of the converted speech was degraded by over-smoothing.

The most recent approaches (He *et al.*, 2002; 2005) use non-numerical clustering techniques to make the conversion. An example of these systems is Uriz *et al.* (2009), which uses a clustering algorithm based on k-histograms (KH): a non-numerical clustering algorithm presented by He *et al.* (2005). These systems have a better performance, with no conditions about a specific distribution for the data. Thus, the cluster is adjusted to the data distribution.

Nevertheless, the problem of creating high-quality voice conversion systems that could be used in real-life applications has not been completely solved. At present, there still is a trade-off between the similarity of converted voices to target voices, and the quality achieved by different conversion methods. The best way to overcome this trade-off is to convert both the vocal tract features of the speaker, and the residual signal of the phonation.

This problem has been faced in other works (Chen *et al.*, 2003; Kain, 2001), where the research was focused on increasing the resolution of GMM-based systems through residual prediction (Duxans and Bonafonte, 2006; Hanzlicek, 2006; Sundermann *et al.*, 2005) in order to improve both the quality scores and the converted-to-target similarity.

This paper proposes a voice conversion (VC) system that combines clustering by means of K-Histograms to convert the vocal tract features with an averaging of residual signals obtained from a pre-recorded dataset for converting the excitation of the voice. This is made to improve both the quality scores and the converted-to-target similarity.

This paper is organized as follows. In Section II, the most important aspects of the voice conversion techniques used in the work are explained in detail. In Section III, a new voice conversion method is proposed. In Section IV, the results of the objective and subjective tests are presented and discussed. Finally, the main conclusions are summarized in Section V.

II. VOICE CONVERSION

The goal of voice conversion systems is to convert the voice of a source speaker, so it is perceived as being pronounced by another specific speaker, who is called the target speaker. The next subsections describe the most important aspects of voice conversion systems.

A. Source-Filter Model

The **source-filter model** (Huang *et al.*, 2001) is a representation of the phonatory system as a filter being excited by a source signal. The filter that represents the vocal tract of the speaker is modeled by using a series of co-axial tubes. This is made by using an *n* degree polynomial, where n is the number of tubes used in the model. The coefficients of this polynomial are called Linear Predictive Coding (LPC). The source of the system is