HAL
open science

Synthetic voice forgery in the forensic context: a short tutorial

Guillaume Galou, Gérard Chollet

To cite this version:

Guillaume Galou, Gérard Chollet. Synthetic voice forgery in the forensic context: a short tutorial. Forensic Speech and Audio Analysis Working Group (ENFSI-FSAAWG), 2011, Rome, Italy. hal-00625918

HAL Id: hal-00625918

https://imt.hal.science/hal-00625918v1

Submitted on 22 Sep 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Synthetic Voice Forgery in the Forensic Context: a short tutorial

Guillaume GALOU
IRCGN
1 bld. Théophile SUEUR 93110 ROSNY-SOUS-BOIS, FRANCE
Gérard CHOLLET
CNRS-LTCI Télécom Paristech, dép. TSI
37-39 rue Dareau 75014 PARIS

***Abstract*—Technical voice forgery in the forensic area has led to several studies, mainly dealing with voice conversion. In the last decade, latests developments around voice synthesis have reached satisfactory intelligibility and quality levels. Moreover several web-based or standalone apps can be used for TTS[1]. Nowadays, HMM-based synthetic voices can be built to fool biometric systems. Several authors reported FARs[2] as high as 70 to 80% when TTS voices where used. Nevertheless, the forensic context is quite different since the human ear might be able to detect a synthetic voice, thus leading to a case discarding. We used the MARY TTS platform in order to produce a speaker-dependent synthetic target voice sample. Given a single synthetic trial, our very preliminary work is to show how feasible and successful such an attack is. Further work is needed to build other voices and provide accurate statistics. Our aim is to confirm a criminal opportunity and to show that synthetic voice detection might become mandatory in a near future.**

***Index Terms*—TTS, speech synthesis, forensic, speaker recognition, voice comparison, HMM, HTS, MARY Text-to-Speech, voice forgery, voice disguise.**

I. Introduction

VOICE forgery is not only a threat for biometrics, but also a threat for identity inference in the forensic context. The performance of commercial forensic speaker verification systems is usually assessed through a standard evaluation protocol like NIST SRE. This is not fully satisfactory from a forensic point of view since impostors attacks are random[6]; robustness to intentional impersonation is not assessed in this case. Whatever the technique, we think that voice forgery is absolutely necessary to be considered when measuring performance of a forensic system, and that natural/transformed/synthesized classifiers front-ends are required.

Facing Forensic Automatic Speaker Recognition (FASR) systems, the aim of a forger is to "provide" either:

- samples from the target (using unit selection synthesis for example) ;
- spectral target-shaped samples (voice transformation or concatenation synthesis).

Keeping in mind that prosodic features are very difficult to forge while preserving quality, most commercial FASR systems do not take into account any prosodic information to model a speaker. But, unlike "voice vaults" applications, forensic speaker recognition can be considered as semi-automated. The forger must also fool the human ear so that:

- any transformation/synthesis cannot be detected;
- perceptual proximity to target must be ensured.

Thus quality and intelligibility are as important as spectral proximity to target. Moreover, considering technical and intentional voice forgery only, good quality and proximity require data samples from the target. Most studies deal with voice conversion [15], [7], [5]. In such a technique, the aim is to estimate a source-to-target acoustic features space transform. Despite a very useful realtime conversion capability, the major drawback is the need of a parallel corpus for training. Such a corpus might not be easy to build and must ensure a full feature space coverage.

As an alternative, unit selection voice synthesis can be used. Although it allows best speech quality and proximity, it is very unlikely since a huge target database is needed. Concatenation synthesis, has also been studied [14], [11].

We believe that Hidden Markov Model synthesis[8], as a special case of concatenation synthesis, is a good candidate for several reasons:

- satisfactory quality vs. affordable target database;
- several toolkits, even web-based TTS, are available;
- prosody can be modeled;
- speaker-adaptation voice modeling is supported (will be discussed further).

Of course, the quality of HMM-based TTS is not perceptually perfect, but it can be considered as satisfactory on short utterances with little background noise.

In our work, we used MARY TTS[13]. It provides a general and handy framework that could be used for criminal purposes in a near future.

II. HMM-based forgery and Automatic Speaker Recognition: previous work

Advances in HMM-based speech synthesis made possible smooth and natural sounding speech. However, until recently, modeling a target speaker required a large amount of data. Now, state-of-the-art HMM synthesis allows model adaptation of speaker-independent models using a small target database [3]. In other words, a first speaker-independent training ensures

[1]Text-to-Speech
[2]False Acceptance (or Alarm) Rate

a broad phonetic coverage and model building whereas only a small speaker-dependent (target) is needed afterwards [10]. Direct speaker-dependent is possible, but phonetically balanced sentences are required. This is the most straightforward strategy. Nevertheless, it is difficult to build such a database using open-source data (web, media, ...).

The authors of [12] developed a text-dependent speaker recognition system. A simple GMM-based approach was used for speaker verification. Given a baseline 0.5 % FAR at EER[3] with natural speech, the system reached 86.3% FAR with synthetic speech at the natural case EER threshold. The EER reached 27% with synthesized voices.

Among the very first works on the subject (1999), the authors of [9] studied a HMM-based text-dependent speaker verification system. From a baseline 0% FAR at EER, the FAR reached between 65% and over 80%, depending on different conditions.

In [4][3], a more recent work, the FAR of GMM-UBM system raised from 0.4% (natural speech) to 90% (synthetic speech at natural EER threshold).

Interest in synthetic voice forgery has grown due to major advances in TTS systems. The results above show that the next big challenge is to provide good quality at an affordable data collection cost. Adaptation allowed drastic training data requirement reduction.

III. HMM-BASED SYNTHESIS

The HMM Speech Synthesis System, H Triple S (HTS) is widely used. Many voices were created in different languages [17]. This system is highly flexible and is promising because its ability to model different styles of speech [18]. As shown in figure1, the system is divided into training and synthesis part.

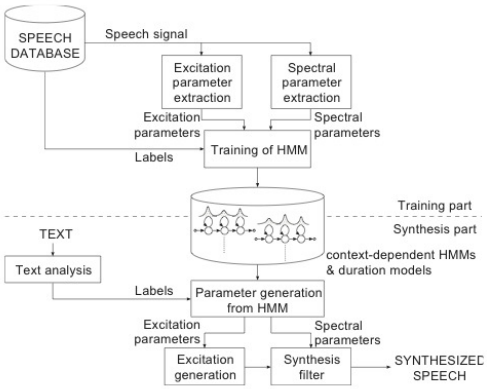

Figure 1. HTS overview (from [18])

The training part is very similar to that used in speech recognition HMM systems (HTS is a patch of HTK). Spectral and source features are used to model context-dependent HMMs; spectrum, excitation and durations are unified in a common HMM framework. Input parameters can be derived from different analysis schemes: MFCCs, PLPs, STRAIGHT, HNM,... Context features are much more complex than in standard speech recognition (mainly based on triphones: single backward and forward phone contexts). To reach a better prosodic modelling, the different dependencies are segments, syllables, words, utterances, sentences, and part-of-speech. Thus, it is necessary to have a phonetizer and a tokenizer. These Natural Language Processing (NLP) tools are not provided with the system.

Waveforms can be synthesized using mel log spectral approximation (MLSA) filter.

As mentioned above, models can be adapted from a previous training. This is a key feature, reducing the needed amount of data for training. In most cases, voice characteristics, speaking styles, or even emotions can be adapted.

IV. IMPACT ON A COMMERCIAL FASR SYSTEM

In our experiment, we used the MARY Text-to-Speech system, an open-source, multilingual Text-to-Speech Synthesis platform written in Java[13]. It was originally developed as a collaborative project of DFKI's Language Technology lab and the Institute of Phonetics at Saarland University and is now being maintained by DFKI.

The reasons for such a choice are:

- french support in Mary TTS has recently been achieved for the first time (french NLP tools and voice) [16];
- Mary TTS provides a handy user-interface and is based on HTS for HMM-based voice training and synthesis;
- Mary architecture is client-server based, thus allowing straightforward and flexible TTS.

For the time being, the latest release does not support up-to-date HTS capability. Especially, speaker adaptation is not supported yet. For the training, sentences from the PolyVar corpus[2] were chosen and recorded in an anechoic chamber. Tokenization and phonetization was realized using LIA PHON[1].

We built a 80-sec unquestioned sample from the recordings. Then both natural and synthetic questioned samples were built (40 sec). 10 listeners were chosen to evaluate both intelligibility and naturalness of the synthetic voice. Mean opinion scores were 4.15/5 and 2.75/5.

The FASR system is BATVOX 3.0. Speaker verification results and samples characteristics are given in table I.

question recording	SNR	LR
natural	25,6 dB	1735
synthetic	43,9 dB	73000

Table I
SPEAKER VERIFICATION RESULTS AND SAMPLES CHARACTERISTICS

Even if these results are not statiscally significant to conclude about the robustness of the system, results in table I are not surprising, even for a state-of-the-art FASR tool. More surprising is the fact that the LR is even larger with the synthetic sample.

[3]Equal Error Rate

V. Conclusions

Unsurprisingly, we confirmed that HMM-based voice forgery is able to fool a state-of-the-art FASR system. Because of its flexibility and its speaker-adaptation capability, HMM synthesis is a good candidate for criminal purposes. High-level applications are already available. Thus, it is quite easy to build either "anonymous" or target voices with small corpora. As a consequence, synthetic voice detection becomes mandatory in the forensic context. Further work is needed in order to produce statiscally significant results.

References

[1] F. Béchet. Lia phon: un système complet de phonétisation de textes. *Traitement automatique des langues*, 42(1):47–67, 2001.

[2] G. Chollet, J.L. Cochard, A. Constantinescu, C. Jaboulet, and P. Langlais. Swiss french polyphone and polyvar: telephone speech databases to model inter-and intra-speaker variability, 1996.

[3] P.L De Leon, I Hernaez, I Saratxaga, M Pucher, and J Yamagishi. Detection of synthetic speech for the problem of imposture. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4844–4847, 2011.

[4] PL De Leon, M Pucher, and J Yamagishi. Evaluation of the vulnerability of speaker verification to synthetic speech. In *Odyssey 2010 - The Speaker and Language Recognition Workshop*, 2010.

[5] M Farrús, D Erro, and J Hernando. Speaker Recognition Robustness to Voice Conversion. 2010.

[6] D. Genoud and G. Chollet. Speech pre-processing against intentional imposture in speaker recognition. In *Fifth International Conference on Spoken Language Processing*, 1998.

[7] Qin Jin, A.R. Toth, A.W. Black, and T. Schultz. Is voice transformation a threat to speaker identification? In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4845 –4848, 31 2008-april 4 2008.

[8] T Masuko. *HMM-Based speech synthesis and its applications*. PhD thesis, Tokyo Institute of Technology, 2002.

[9] T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi. On the security of hmm-based speaker verification systems against imposture using synthetic speech. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 3, pages 1223–1226, 1999.

[10] S Oller, A Moreno, and A Bonafonte. Synthesis using speaker adaptation from speech recognition db. In *FALA-2010*. ISCA Special Interest Group on Iberian Languages (SIG-IL), Multimedia Technology Group (GTM), Spanish Thematic Network on Speech Technology (RTTH), 2010.

[11] P. Perrot, G. Aversano, and G. Chollet. Voice disguise and automatic detection: review and perspectives. *Progress in nonlinear speech processing*, pages 101–117, 2007.

[12] T Satoh, T Masuko, and T Kobayashi. A robust speaker verification system against imposture using an hmm-based speech synthesis system. In *EUROSPEECH-2001*, volume 759-762, 2001.

[13] Marc Schröder and Jürgen Trouvain. The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology*, 6(4):365–377, 2003.

[14] Y Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *Speech and Audio Processing, IEEE Transactions on*, 9(1):21–29, 2002.

[15] Y Stylianou, O Cappé, and E Moulines. Continuous probabilistic transform for voice conversion. *Speech and Audio Processing, IEEE Transactions on*, 6(2):131–142, 2002.

[16] F. Xavier. Synthèse vocale - intégration du français au système mary text-to-speech. Master's thesis, UPMC - Telecom Paristech - IRCAM, 2011.

[17] J Yamagishi, B Usabaev, S King, O Watts, J Dines, Jilei Tian, Yong Guan, Rile Hu, K Oura, Yi-Jian Wu, K Tokuda, R Karhila, and M Kurimo. Thousands of Voices for HMM-Based Speech Synthesis–Analysis and Application of TTS Systems Built on Various ASR Corpora. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(5):984–1004, 2010.

[18] H Zen, T Nose, J Yamagishi, S Sako, T Masuko, AW Black, and K Tokuda. The HMM-based speech synthesis system (HTS) version 2.0. *Proc. of Sixth ISCA Workshop on Speech Synthesis*, pages 294–299, 2007.