Athens Institute for Education and Research
ATINER

ATINER's Conference Paper Series
COM2016-2071

Speech Synthesis of Central Mexico Spanish using Hidden Markov Models

Carlos Franco
PhD Student
National Autonomous University of Mexico
Mexico

Abel Herrera Camacho
Professor
National Autonomous University of Mexico
Mexico

Fernando Del Rio Avila
Research Assistant
National Autonomous University of Mexico
Mexico

An Introduction to ATINER's Conference Paper Series

ATINER started to publish this conference papers series in 2012. It includes only the papers submitted for publication after they were presented at one of the conferences organized by our Institute every year. This paper has been peer reviewed by at least two academic members of ATINER.

Dr. Gregory T. Papanikos
President
Athens Institute for Education and Research

This paper should be cited as follows:

Franco, C., Herrera Camacho, A. and Del Rio Avila, F. (2016). **"Speech Synthesis of Central Mexico Spanish using Hidden Markov Models",** Athens: ATINER'S Conference Paper Series, No: **COM2016-2071.**

Athens Institute for Education and Research
8 Valaoritou Street, Kolonaki, 10671 Athens, Greece
Tel: + 30 210 3634210 Fax: + 30 210 3634209 Email: info@atiner.gr URL: www.atiner.gr
URL Conference Papers Series: www.atiner.gr/papers.htm

ISSN: 2241-2891
01/12/2016

Speech Synthesis of Central Mexico Spanish using Hidden Markov Models

Carlos Franco

Abel Herrera Camacho

Fernando Del Rio Avila

Abstract

The current century has proved being relevant in the design of new speech synthesizers. The incorporation of Hidden Markov Models HMM has changed the paradigm in the design of concatenative speech synthesizers. Such systems are called HMM text to speech synthesis (HTS). This paper describes a version adapted to central Mexico Spanish. A MOS test shows an intelligibility score of 3.4 and 3.1 of naturalness.

Keywords: Festival speech synthesis system, Hidden Markov Models, HTS synthesis technique, Speech Synthesis.

Acknowledgments: The authors would like to thank the support of the project PAPIIT IT 102314 de DGAPA-UNAM; to CCADET de of UNAM for the use of their facilities, recording equipment and technical support of Dr. Felipe Orduña Bustamante; to engineer René Ernesto Mendoza Sánchez of Engineering School at UNAM who recorded the voice; to Master of Engineering Carlos Acosta for the logistic support; and the financial support to Posgrado de Ciencia e Ingeniería de la Computación of UNAM to present this work.

Introduction

Text to Speech synthesis aims to get closer to the goal of generating a synthetic voice indistinguishable from that of a real person. The described system´s synthetic voice was primarily recorded from an actual person. The phonemes are treated separately to generate any desired phrase. Therefore, the possibility to generate new or mixed voices is open.

Historically, Speech Synthesis research begins during the second half of the twentieth century. Many changes have taken place during the last 50 years. The predominant synthesizers designed nowadays belong to the concatenative type. They use phonemes or sub-phonemes as units; the units are then parametrized (e.g MFCC), stored and then ordered. They can be sought using deterministic selection trees. Figure 1 shows a block diagram of the different stages in a concatenative text to the speech synthesis system. System input is unrestricted text in the form of a character sequence, including numbers, abbreviations and punctuation signs. The function of the text normalizer is to process any non-alphanumeric characters.

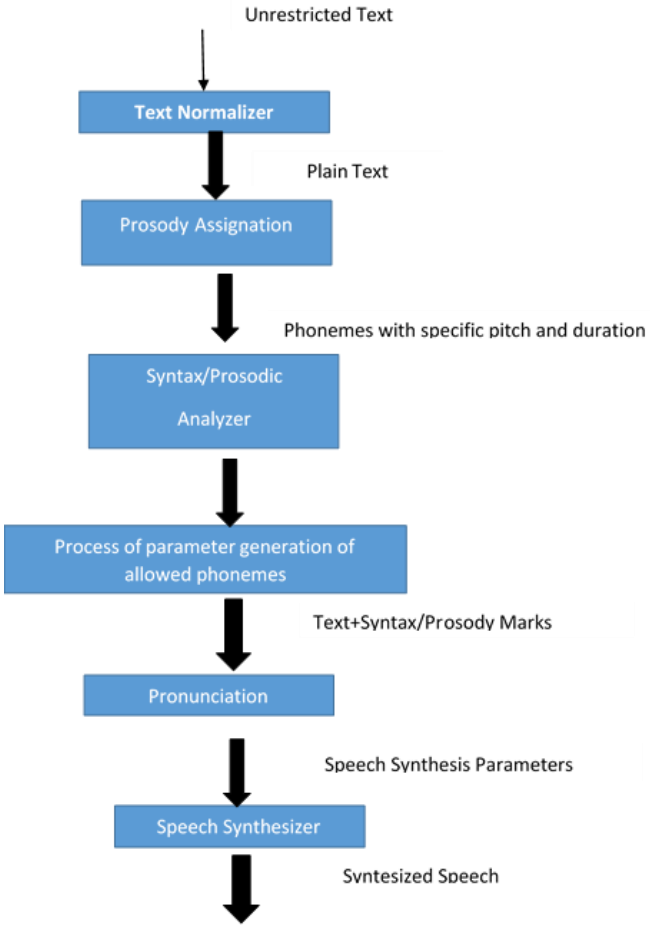
The following stages are basically the syntax/prosody analysis; text is fragmented in such a way that significant intonation and duration is added.

Since the early 80s, CMU together with the University of Edinburgh created Festival. In such a system, they used deterministic trees from a database called CLUNITS (Black and Taylor, 1997). The authors developed a similar system for Central Mexico Spanish (Del Río and Herrera, 2003); the results, from both groups are very intelligible and acceptable natural voices.

Hidden Markov Models (HMM) selection trees were developed by several research groups since the end of the last century (Falaschi et al., 1989; Giustiniani and Pierucci, 1991). However, the most remarkable HMM as text to speech synthesis (HTS) is the work of Dr. Tokuda and his group. (Tokuda et al., 2000; Tokuda et al., 2002; Heiga and Takashi, 2005). The authors developed an HTS system for central Mexico Spanish using non-professional recordings (Herrera and Del Río, 2011).

Also, FESTIVAL produced their own system based on HTS, called CLUSTERGEN (Black, 2006). Several parameterizations have been designed to improve the HTS synthesis quality, STRAIGHT being the most relevant among them (Kawahara et al., 1999; Chang and Kewley-Port, 2004).

Figure 1. *Block Diagram of a Concatenative Synthesizer*



Hidden Markov Models

The Hidden Markov Models basic theory was published by Baum and his colleagues in a series of articles in the late 60s and early 70s. It was implemented by Baker in a voice processing in CMU and by Jelinek and his colleagues in IBM in the 70s. Nevertheless, it was during the second half of the 80s when the HMM theory in speech processing was better understood and applied. Relevant publications are Rabiner´s paper on HMM and his speech recognition book (Rabiner, 1989; Rabiner and Juang, 1993).

A discrete model will be presented to introduce HMMs. An HMM is a set of states {Si}, whose transition probabilities are {aij}, and the observation probabilities of each state are {bj (k)}, each state has initial probabilities {πij}. Figure 2 shows a Markov chain of N different states: S1, S2,…, SN, where N = 5 in terms of simplicity.

In equally spaced time intervals, the system experiments a change of state (it is possible to return to the same state it began) according to a set of probabilities associated to that state. Let t=1, 2, … be the time instants when the state changes and let qt be the state of q at time t. A complete probabilistic description would require to specify the current state (in time t), as well as every previous state. For the special case of a discrete Markov

chain, this description is simplified to the current state and its predecessor. This is described in equation (1).

$$P\left[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \ldots\right] = P\left[q_t = S_j | q_{t-1} = S_i\right] \quad (1)$$

Besides, the only considered processes are those to the right of the equation (1) which are time independent. Therefore, the system has a set of transition probabilities in states $\mathbf{a_{ij}}$ as indicated in equation (2).

$$a_{ij} = P\left[q_t = S_j | q_{t-1} = S_i\right] \quad 1 \leq i, j \leq N \quad (2)$$

Where the states transition probabilities have basic probability properties described in equation (3), they conform a matrix A.

$$a_{ij} \geq 0 \quad (3a)$$

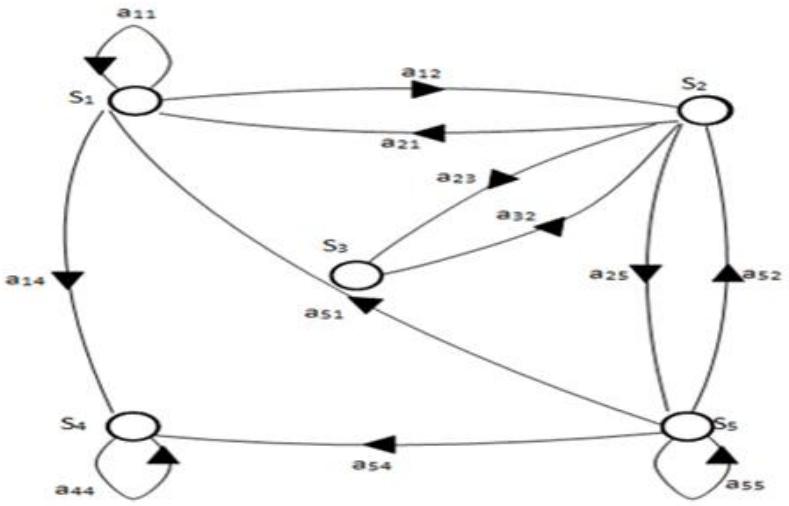
$$\sum_{j=1}^{N} a_{ij} = 1 \quad (3b)$$

Initial state probabilities are defined in equation (4):

$$\pi_i = P[q_1 = S_i] \quad 1 \leq i \leq N \quad (4)$$

The probability distribution function of the symbol observed in j state, B = {b_1 (k)}, is expressed in equation (5), where M is the number of observations in a state. All states have the same number of observations and the same observations.

Figure 2. *HMM with Five States*



$$b_j(k) = P\left[v_k \text{ in } t \mid q_t = S_j\right] \quad 1 \le j \le N \quad 1 \le k \le M \quad (5)$$

From the argument above exposed, it should be noticed that the complete specification of a HMM requires the specification of two parameters of the model (N and M), the specification of the observed symbols and the specification of three probabilistic measures A, B and π. For the sake of simplicity, it is used the notation expressed in equation (6) to indicate the complete set of the model parameters.

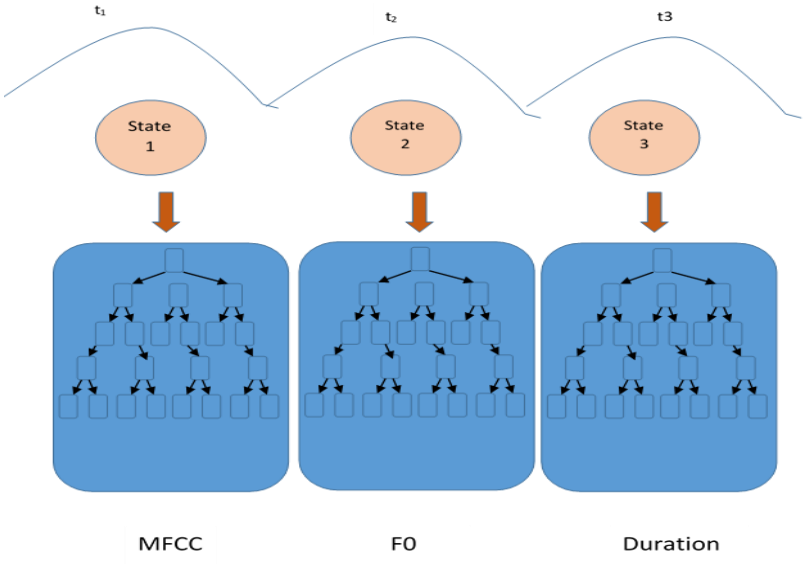
$$\lambda = (A, B, \pi) \quad (6)$$

Given the form of HMM, there are two basic problems for voice processing that need to be solved. Both problems are related on How to train the HMM (problem 1) and how to classify a phrase efficiently calculating $P(O|\lambda)$, given the observation sequence $O = O_1O_2\ldots O_T$ and model $\lambda = (A, B,\pi)$ (problem 2).

In HMMs there is a third problem which has little to do with speech processing. It consists of how to select a Markov model which most resembles a test phase (problem 3). Problem 1 is about optimizing the model parameters the best possible way, by studying a given observation sequence. The observation sequence used to adjust the model parameters is called training sequence, since it is used to train the HMM. The training problem is crucial for most HMM applications, since it efficiently adapts the model parameters to the information observed during the training stage, which means, creating the best models to real phenomena.

Problem 2 is where the hidden part of the model is to be discovered, it means to find the correct sequence of states. It has to be clear for all cases, except degenerative models; there is no correct sequence of states to find. Therefore, for practical situations, some optimization criteria are applied to solve the problem accurately.

Figure 3. *Context Dependent HMM with Vectors*



Unfortunately, there are several optimization criteria to choose from, therefore, eligibility, is a strong aspect in the task mathematical solutions to HMM, but the criteria are not presented; the references can be looked upon for that matter.

HTS Synthesis

A speech signal is decomposed in three groups of data: Excitation F0, Mel Cepstral Frequency Coefficients and duration.

Each of these three groups is labeled and separately stored in Gaussian vectors. Each of them is at the same time a state vector in the HMM. This is done in order to lay up multiple variants of the same phonemes, see Figure 3. The reason behind it is to provide the system with enough options to produce natural sounding utterances. State durations of each HMM are modeled by a multivariate Gaussian distribution. The dimensionality of a state duration density corresponds to the number of states in the HMM. The data contained in such trees will be selected afterwards using Context dependent HMMs.

Context Dependent HMMs are needed to assemble the best choice of each F0, MFCC and density duration contained. They are determined by maximizing their output probability.

The selection, once assembled is synthesized by adjusting the received parameters in a MLSA filter which is fed by a sinusoidal source for voiced sounds and a noise source for unvoiced sounds. Figure 4 summarizes the process. A detailed description of the MFCCs, F0, duration as well as the context duration HMMs is presented.

MFCCs used on MLSA Filter

At this point, much has been said about Mel Frequency Cepstral Coefficients for speech parameterization. In this case the MFCC conform a vector in the HMM. The MFCC will determine the parameter used in the Mel log Spectrum approximation filter which will produce the synthesized speech (Tokuda et al., 2002).

F0

Two types of sounds can be found during speech: voiced and unvoiced sounds (e.g. vowels and fricatives respectively). Voiced sounds clearly show a periodic F0 frequency, whereas unvoiced sounds lack of periodicity. These can be interpreted as noise.

Determination of F0 for voice segments has been a challenge since the 1960s. There are several methods aiming for that goal without succeeding in terms of precision for all possible phonemes in continuous speech. The algorithm proposed by Goncharoff and Gries (1998) has been selected for the current work.

The F0 data is also stored in a vector which conforms another state in the HMM. The voiced sounds are represented by a continuous vector whereas the unvoiced sounds are conformed by a discrete vector. Therefore, a mixture of continuous/discrete HMM is proposed by Tokuda and colleagues (Tokuda et al., 2002).

Context Dependent HMMs

It would be impossible to prepare a speech database including all combinations of contextual factors, such as: phone identity, stress relation and locality. Thus, context dependent HMM were used.

When adapted to english by Tokuda et al., (2002), the contextual factors of the HMM follows the following enquiries: Utterance to phrase, phrase to word, word to syllable and syllable to phoneme. The same contextual factors were taken into account when the system was adapted to Spanish.

Phoneme:

- (preceding, current, succeeding) phoneme
- position of current phoneme in current syllable

Syllable:

- number of phonemes at (preceding, current, succeeding) syllable
- accent of (preceding, current, succeeding) syllable
- stress of (preceding, current, succeeding) syllable
- position of current syllable in current word number of (preceding, succeeding) stressed syllables in current phrase.
- number of (preceding, succeeding) accented syllables in current
- phrase
- number of syllables (from previous, to next) stressed syllable
- number of syllables (from previous, to next) accented syllable
- vowel within current syllable

Word:

- guess at part of speech of (preceding, current, succeeding) word
- number of syllables in (preceding, current, succeeding) word
- position of current word in current phrase
- number of (preceding, succeeding) content words in current phrase
- number of words (from previous, to next) content word

Phrase:

- number of syllables in (preceding, current, succeeding) phrase
- position in major phrase
- end tone of current phrase

Utterance:

- number of syllables in current utterance

The procedure was done using Festival functions for speech labeling focused on Spanish grammatical rules via a piece of code named *lexicon.*

The input voice signal to train the system consisted of 300 phrases recorded as wave audio files. A phonetically balanced text was recorded with a duration of 20 minutes. The recording took place at an anechoic chamber performed by a professional radio host. The signal was sampled at 16 kHz. The sentences contents are from general topics. The synthetic speech timbre was remarkably like that of the speaker.

Evaluation of System

A Mean Opinion Score (MOS) test was applied to 40 people, 30 men and 10 women. The average age of the group was 18 to 25 years old. No one had suffered an auditory system illness.

The test was a questionnaire where the subject had to listen to an audio file which consisted of a phrase recorded by a human speaker followed by the same utterance produced by the HMM synthesizer. There was a two second pause of silence between both phrases.

Each subject listened to five pair of phrases, the subject marked a grade for each synthetic phrase.

The evaluated factors were naturalness and intelligibility. The listener valued in a scale of 0 to 5, being 0 the less natural/intelligible and 5 being the best. The test took around 3 minutes per listener. The total average scores found were 3.6 in naturalness and 3.4 in intelligibility. Figure 5 shows the graphic results, the orange lines represent intelligibility, whereas the blue ones stand for naturalness. Each line represents the average of the sentences for that listener.

The MOS test clearly show that the quality of our synthesized phrases is beyond the mean but according to the listener’s opinions we still need some improvement to reach the maximum marks in both: intelligibility and naturalness.

Figure 4. *Speech Synthesis Scheme*

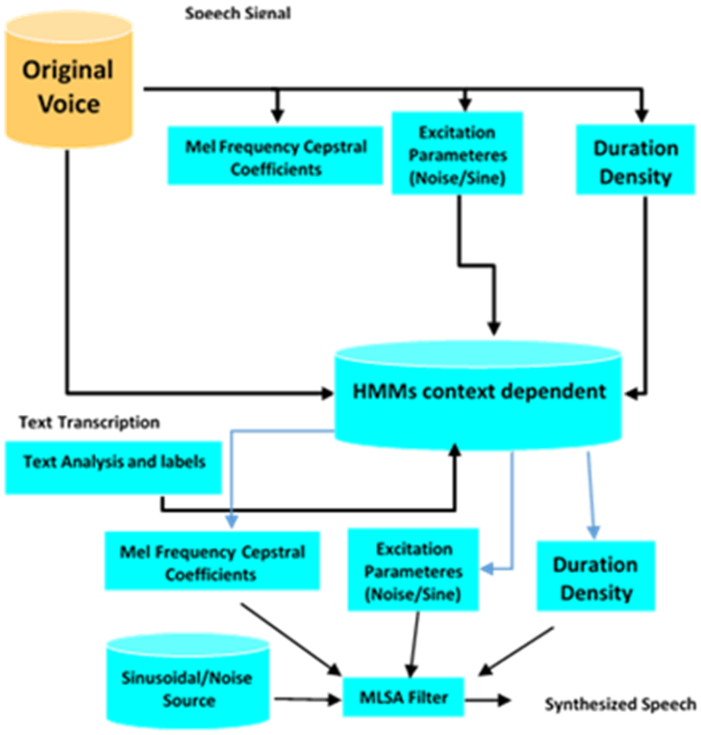
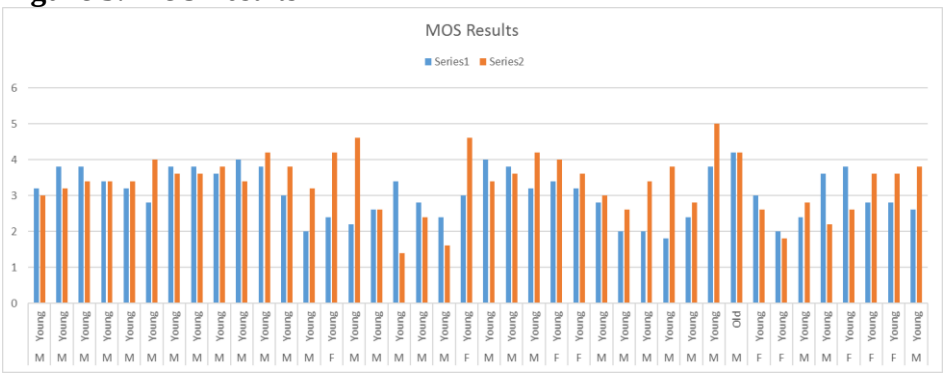


Figure 5. *MOS Results*



Conclusions and Future Work

A proposed method to improve our system consists on trying out a parameterization method alternative to the MFCCs which is currently the most popular for speech. The authors considered using Linear Spectral Pairs LSP which is directly derived from the classic Linear Prediction Coding voice parameterization. As reported by (Heiga et al., 2008) LSP based synthesis probed to be high quality.

The impact of the current synthesis research is palpable since very few attempts of speech synthesis in the Spanish language exist in the literature. A successful system was developed in the Universidad de Bilbao (Erro et al., 2014) but their aim was basically the Basque language.

The described system is currently one of the fewest focused on central Mexico Spanish, the reason for using HMMs relies on the fact that the

system can virtually "say" any written phrase. Whereas other speech synthesizers, e.g. unit-based ones, are always limited to their corpus size.

Another important challenge is to give the system changes in expression, at this point, even when the voice naturalness has improved the phrases lack the expression to resemble a human being.

References

Black, Alan W and Taylor, Paul (1997). "Automatically Clustering Similar Units for Unit Selection in Speech Synthesis", http://bit.ly/2gEgFqJ.

Black, A. W. (2006). "CLUSTERGEN: a statistical parametric synthesizer using trajectory modeling." Proceedings of INTERSPEECH 2006.

Chang, Liu and Kewley-Port (2004). "STRAIGHT: A new speech synthesizer for vowel formant discrimination". Acoustic Research Letters Online, volumen 5, número 31.

Del Río, F. and Herrera, A. (2003) "A Mexican Speech Concatenative Synthesis System by Diphones" . Memorias del 10th Congreso Mexicano de Acústica.

Erro, D., Sainz, I., Navas, E. and Hernaez, I. (2014). Harmonics plus noise model based vocoder for statistical parametric speech synthesis. IEEE Journal of Selected Topics in Signal Processing, 8(2), 184-194.

Falaschi, A., Giustiniani, M. and Verola M. (1989). "A hidden Markov model approach to speech synthesis." Proceedings of EUROSPEECH 1989, pp. 187-190.

Giustiniani, M. and Pierucci, P. (1991). "Phonetic Ergodic HMM for speech synthesis." Proceedings of EUROSPEECH 1991, pp. 349-352.

Goncharoff, V. and Gries, P. (1998). " An algorithm for accurately marking pitch pulses in speech signals", Proceedings of the IASTED International Conference on Signal and Image Processing.

Heiga, Zen1, Takashi, Nose (2005). "The HMM-based Speech Synthesis System (HTS) Version 2.0", http://bit.ly/2gElf8C.

Heiga, Z. E. N., Tomoki, T. O. D. A. and Tokuda, K. (2008). The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006. IEICE transactions on information and systems, 91(6), 1764-1773.

Herrera, A. and Del Río, F. (2011) "Development of a Mexican Spanish HMM-Based Synthetic Voice". Memorias del 18th Congreso Mexicano de Acústica.

Kawahara, H., Masuda-Kastuse, I., and Cheveigne, A. (1999). ''Restructuring speech representations using a pitchadaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds''. Journal of Speech Communication, número 27, pp 187-207.

Rabiner, L. R. (1989). "A Tutorial of Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, vol. 77 (2), pp. 257-286.

Rabiner, L. R., and B. H. Juang (1993), Fundamentals of Speech Recognition. Englewood Cliffs, Prentice Hall.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T. (2000). "Speech parameter generation algorithms for HMM-based speech synthesis." ". Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 1315-1318.

Tokuda, K., Zen, H. and Black, A.W. (2002) "An HMM-based speech synthesis system applied to English". Proceedings of IEEE SSW.