ارائـــه روش[1] TD-VQSS برای سنتـــز گفتـــار سیستم‌های[2] TTS

ابوالقاسم صیّادیان
دانشیار
دانشکده مهندسی برق، دانشگاه صنعتی امیرکبیر

چکیده

هرسیستم TTS از دو قسمت اساسی؛ الف) قسمت LA[3] یا NLPP[4] ب) قسمت SS[5] تشکیل شده است. وظیفه قسمت LA تشخیص و تعیین نوع جمله (سؤالی، خبری، ...) تعیین موقعیت و نقش معنائی یا دستوری کلمه در جمله، تعیین زنجیره آوائی (یا سیلابی) کلمات همراه با نوع لحن و تأکید بیان و ... می باشد. وظیفه قسمت SS تولید سیگنال صحبت از روی زنجیره آوائی (یا سیلابی) خروجی قسمت LA و اعمال لحن و تأکید بیان مطلوب بر روی آنها می باشد. تحقیق این نوشتار بر روی قسمت SS متمرکز می باشد. پس از معرفی روش‌های کلاسیك مطرح برای قسمت SS، روش جدیدی به نام TD-VQSS برای این منظور ارائه می شود. این روش دارای ویژگی‌های زیر می باشد: نیاز به حجم حافظه کم ـ نیاز به بار محاسباتی پایین ـ تولید صدای کاملاً طبیعی و قابل فهم ـ امکان تغییر لحن و تأکید بیان گفتار و ...

TD-VQSS a New Speech Synthesizer for TTS Systems

A. Sayadian.
Associate Professor
Electerical Engineering Department,
Amirkabir University of Technology

Abstract

Every TTS system has two principal sections:
a) Language Analyzer (LA) or Natural Language parser (NLP)
b) Speech synthesizer (SS). The tasks of LA section are determination of: sentence type, phoneme sequence of words, prosodic contour of phonemes .... . The task of SS section is production of speech signal from phonemes sequences with proper prosody. The research of this paper is concentrated on SS section. After survey of existing method, we propose a new speech synthesizer with the name of TD-VQSS. The features of this new method are: Very highna aturalness speech production, possibility of prosodic variation, low storage, low computational load.

کلمات کلیدی

Text-to- Speech, Speech Synthesizer, Pitch Synchronous Overlap Add (PSOLA)

مقدمه

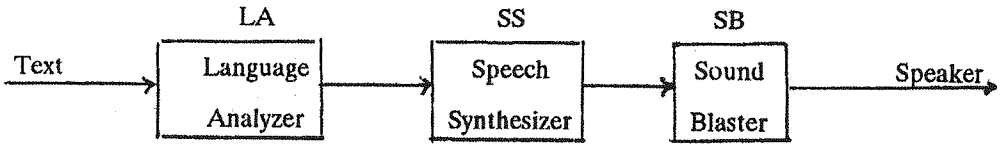
ارتباط گفتاری بین انسان و ماشین (کامپیوتر، ربات و ...) مشتمل بر دو سرفصل اساسی در پردازش گفتار و پردازش زبان طبیعی می باشد. سرفصل اول شامل روش های تبدیل گفتار به متن و سرفصل دوم شامل روش های تبدیل متن به گفتار (به یک یا چند زبان مشخص) می باشد. زمینه تحقیقات این نوشتار مربوط به سرفصل دوم یعنی سیستم های تبدیل متن به گفتار (TTS) است. ورودی هر سیستم TTS نوعاً یک جمله نوشتاری [۶] یا یک فایل نوشتاری [۷] است (که می تواند مشتمل بر دهها یا صدها جمله باشد). مجموعه وظایف هر سیستم TTS به دو قسمت اساسی به نام های LA و SS سپرده می شود [۴] ، ... ، [۱].

وظیفه قسمت LA، تشخیص نوع جمله (خبری، تعجبی، امری، سؤالی و ...) تعیین نقش معنائی و دستوری کلمات در جمله، تعیین زنجیره آوائی (یا سیلابی) کلمات جمله، تعیین نوع لحن و تأکید بیان (در نتیجه تعیین نوع الگوی کانتور تغییرات فرکانس پیچ و گین و نسبت های طول واجها و سکوت و ...) و .. می باشد. وظیفه قسمت SS تولید سیگنال صحبت حتی الامکان با لحن طبیعی و با نوع تأکید بیان مطلوب و صدای گوینده مورد نظر (زن یا مرد) می باشد. تحقیقات اصلی این نوشتار بر روی قسمت SS سیستم های TTS متمرکز می باشد.

در بخش دوم روش های موجود سنتی سایزر گفتار را مورد بررسی و تحلیل مختصر قرار می دهیم. در بخش سوم به معرفی و تحلیل روش جدید TD-VQSS می پردازیم. سرانجام در بخش چهارم جمع بندی از نتایج تحقیقات این نوشتار خواهیم داشت.

۲ ـ روش های کلاسیک در سنتز گفتار

اگر سنتز گفتار برای تعداد کلمات یا جملات محدودی مورد نظر باشد، ساده ترین روش استفاده از ضبط کلمات یا جملات به صورت کامل از یک یا چند گوینده است. این روش اگرچه در تعداد محدودی از کاربردها استفاده می شود، ولی انعطاف پذیری لازم برای اغلب کاربردها را ندارد. در سیستم های TTS که برای تولید نامحدود کلمات یا جملات یک زبان طراحی می شوند، ضرورت دارد که از واحدهای کوچکتر زبانی مانند سیلاب [۸]، نیمه سیلاب [۹]، واج ترکیبی[۱۰] یا واج [۱۱] استفاده شود، [۸] ... [۵]. هر قدر واحد زبانی کوچکتر انتخاب گردد، تعداد آنها کمتر (مانند واج)، ولی مکانیزم های کنترلی برای اتصال آنها [۱۲] پیچیده تر و امکان تولید صدای کاملاً طبیعی مشکل تر خواهد بود. برعکس هر قدر واحد زبانی بزرگتر باشد (مثل سیلاب)، صدای تولید شده طبیعی تر و مکانیزم های کنترلی ساده تر شده و در عوض حجم حافظه مورد نیاز برای نگهداری آنها بیشتر می گردد. پس از انتخاب نوع واحد زبانی، مسئله بعدی انتخاب نوع سنتز کننده سیگنال گفتار در سطح فریم (ms ۲۰ –۵) می باشد. اگر سنتزکننده گفتار برای تولید نامحدود کلمات یک زبان به کار رود، با شاخص های زیر مورد ارزیابی قرار می گیرد: الف) امکان سنتز صدای قابل فهم، ب) امکان تغییر فرکانس پیچ و گین (برای تغییر لحن و نوع تأکید بیان)، ج) امکان سنتز صدای طبیعی، د) امکان تشخیص گوینده، ذ) حجم حافظه مورد نیاز ر) بار محاسباتی سنتزکننده. روش هایی که از سنتز در حوزه زمان استفاده می نمایند (مانند روش[۱۳] PSOLA [۱۳]، ...، [۱۰] از صدای طبیعی و قابل فهم بسیار بالایی برخوردار بوده ولی در عوض به حجم حافظه نسبتاً بالایی نیاز دارند. روش هایی که از سنتز پارامتریک غیرزمانی (مانند سنتی سایزر[۱۴] فرمنت کلات [۹] [۶] [۵] و یا سنتی سایزر[۱۵] LSPF [۷] استفاده می نمایند، صدای قابل فهم تولید نموده ولی در تولید صدای طبیعی دچار مشکل بوده و علاوه بر آن دارای بار محاسباتی



شکل (۱) شمای کلی یک سیستم TTS.

بالائی در هنگام سنتز می باشند. درعوض حجم حافظه مورد نیاز برای نگهداری پارامترها نسبتاً پایین است (در مقایسه با سنتز کننده های حوزه زمان). آنچه تاکنون در سیستم های حرفه ای مورد استفاده قرار گرفته، روش های پارامتریک غیرزمانی می باشد (مانند سیستم ME8000, Votrax SCO1A Digital DTC-05, General Instrument SP256-ALZ, NEC-PD7752C, Silicon Systems SSI263, Philips Electronic Dectalk(و ... [۱۵] ، [۱۴]. این نوع سنتی سایزرها به علت نیاز به حجم محاسبات بالا، نوعاً توسط یک یا دو چیپ ارائه می شوند. با توجه به اینکه هدف عمده این تحقیق آن است که بار محاسباتی قسمت سنتی سایزر کم بوده، به قسمی که بر روی کامپیوترهای PC (بدون استفاده از چیپ های مخصوص) به راحتی قابل پیاده سازی باشد، به نظر می رسد که روش های زمانی کاندید مناسبی برای اهداف فوق می باشند. عمده ترین روش زمانی مطرح که قادر به تغییر لحن و تأکید بیان نیز هست روش PSOLA می باشد. با اینکه روش فوق قادر به تولید صدای کاملاً طبیعی و قابل فهم با بار محاسباتی کم می باشد، متأسفانه به علت نیاز به حجم حافظه بالا، هنوز وارد سیستم های حرفه ای نشده و در مراکز تحقیقاتی و دانشگاهی مورد استفاده قرار می گیرد. دلیل اصلی نیاز به حجم حافظه بالا این است که، در این روش کل قطعات زبانی (واج، واج ترکیبی یا ...) را یکجا و به صورت زمانی (با فرکانس ۱۱/۰۲۵KHZ-۸ و به صورت ۸ یا ۱۶ بیتی) ذخیره می نمایند. آنچه تحقیق این نوشتاررا به خود معطوف نموده، ارائه روشی است که ضمن دارا بودن محاسن روش PSOLA، به حجم حافظه کمتری نیاز داشته به قسمی که برای سیستم های TTS با تعداد کلمات نامحدودی قابل استفاده باشد.

۳ ـ ارائه روش TD-VQSS

توانایی روش ارائه شده در کاهش حجم حافظه مورد نیاز قسمت SS سیستم های TTS می باشد. این روش از سه نوع تئوری و مدل سازی به شرح زیر استفاده می نماید: الف) آشکارسازی نقطه تحریک [۱۶] برای فریم های باصدا[۱۷] [۱۶] ب) چندی کننده برداری[۱۸] VQ برای کاهش نرخ بیت ج) درون یابی[۱۹] و چند به یک کردن[۲۰] سیگنال در حوزه زمان.

۱ ـ ۳ ـ آشکارسازی نقطه تحریک برای فریم های باصدا

در تئوری تولید سیگنال صحبت، هر فریم زمانی (۲۰-۵ ms) را می توان به دو کلاس کلی با صدا و بی صدا تقسیم نمود. اطلاعات فاز و نقطه تحریک در فریم های بی صدا اهمیتی برای گوش انسان ندارند، ولی چنانچه بخواهیم سیگنال گفتار کاملاً طبیعی و قابل فهم تولید نماییم، اطلاعات فاز (نسبی) و نقطه تحریک برای فریم های باصدا بسیار حائز اهمیت می باشند. برای درک اهمیت آشکارسازی صحیح نقطه تحریک در سنتز زمانی، سیگنال صحبت تعدادی از جملات را با اتصال متوالی یک پریود پیچ (با گین های متناظر) برای قسمت های باصدا و اتصال متوالی قطعاتی به طول ۱۰ ms برای قسمت های بی صدا تولید نمودیم. فرض می کنیم در برش هر فریم زمانی، به جای نقطه تحریک از نقطه ای به طول l از نقطه تحریک $0 \le l \le PP/2$ که PP طول پریود پیچ برحسب نمونه های سیگنال می باشد عمل نمائیم. صدای بازسازی شده را مورد آزمون های شنیداری قرار دادیم. نتیجه در جدول (۱) درج شده است. اعداد ردیف اول جدول نسبت l/PP بوده که به صورت درصد بیان شده اند.

جدول (۱) نتیجه آزمون های شنیداری برش فریم های زمانی نسبت به نقطه تحریک.

بایاس نقطه تحریک	٪۰	٪۱۰	٪۲۰	٪۳۰	٪۴۰	٪۵۰
نتیجه آزمون شنیداری	٪۹۸/۵	٪۹۴/۳	۸۲/۴	٪۷۵/۷	٪۷۴/۵	٪۷۲/۶

روش آزمون بدینگونه بوده است که جملات بازسازی شده برای ۱۰ نفر پخش شده است، شنوندگان با مقایسه کیفیت صدای بازسازی شده و صدای اصلی نمره ای بین صفر تا صد به صدای بازسازی شده داده اند. آنچه در جدول (۱) درجه گردیده متوسط نمرات پاسخ ۱۰ نفر می باشد. صدای بازسازی شده برای $l > 0$ دارای حالت گلوگرفتگی و همچنین کلیک های پریودیک می باشد. بنابراین به نظر می رسد که آشکارسازی نقطه تحریک فریم های باصدا برای سنتز زمانی گفتار به صورت اتصال متوالی فریم هایی به طول پریود پیچ، امری ضروری است. روش های مختلفی برای آشکارسازی نقطه تحریک پیشنهاد و مورد استفاده قرار می گیرند. با توجه به اینکه فرآیند آشکارسازی در طی این تحقیق به

صورت زمان زنده[۲۱] مورد نظر نبوده و همچنین دقت آن روش مورد توجه می باشد، از روش مندرج در مرجع [۱۶] استفاده شده است.

۲-۳- استفاده از چندی کننده برداری VQ برای کاهش حجم حافظه

در مبحث مربوط به کد کننده های صحبت با نرخ پایین نشان داده می شود که بهترین راه، جهت وصول به نرخ های خیلی پایین (۲۴۰۰BPS - ۸۰۰) استفاده از چندی کننده های برداری است [۱۸] [۱۷]. برای استفاده از توانایی چندی کننده های برداری در کاهش حجم حافظه مورد نیاز سیستم TTS، به شرح زیر عمل می نماییم. ابتدا پس از ضبط و پالایش و طراحی قطعات زبانی مورد نظر (واج، ترکیب واج، نیمه سیلاب و ..) به هر فریم ۱۰ میلی ثانیه از قطعات فوق الذکر، یک برچسب V یا UV (با صدا یا بی صدا) می زنیم. سپس فرآیندهای زیر را انجام می دهیم:

۱ - برای فریم های با صدا، ابتدا با استفاده از روش مندرج در بخش قبل، محل نقطه تحریک و پریود پیچ را تخمین می زنیم. برای فریم های بی صدا نیازی به تخمین نقطه تحریک و پریود پیچ نداریم.

۲ - قطعه زمانی به اندازه یک پریود پیچ از نقطه تحریک (برای فریم های با صدا) و یا یک قطعه به طول ۱۰ میلی ثانیه از نقطه شروع فریم (برای فریم های بی صدا)، متناظر با کلیه فریم های موجود را به دست آورده و ذخیره می نماییم.

۳ - هر قطعه زمانی را به صورت پریودیک به دنبال هم سه بار تکرار نموده و آنگاه نمونه های زمانی حاصل را در یک پنجره همینگ به طول ۲/۵ برابر پریود قطعه (و یا ۲۰ میلی ثانیه برای فریم های بی صدا) ضرب می نماییم. با استفاده از آنالیز[۲۲] LPC ، ۱۲ ضریب[۲۳] LSPF متناظر را برای هر قطعه زمانی فوق الذکر به دست می آوریم.

۴ - با استفاده از پارامترهای LSPF و روش های طراحی کتاب کد سیستم VQ همانند روش های[۲۴] LBG [۱۷] یا[۲۵] FSL [۱۸]، تعداد ۱۰۲۴ کلمه کد برای قسمت SS سیستمTTS طراحی می نماییم.

۵ - فرض می کنیم $C = \{C_1, C_2, ... C_{1024}\}$ بردارهای مرجع در کتاب کدنهایی باشند، برای هر بردار C_i در کتاب کد C نزدیکترین الگو از میان کلیه بردارهای آموزشی را به دست آورده و سیگنال زمانی متناظر آن را $S_i(n)$ می نامیم.

۶ - با استفاده از روش درون یابی با باند محدود[۲۶] [۲۰] [۱۹]، طول قطعات زمانی متناظر C_i ها را به ماکزیمم طول پریود پیچ عملیاتی یعنی ۱۲/۵ ms می رسانیم (فرآیند مرحله ۶ فقط برای قطعات زمانی که دارای برچسب با صدا می باشند اجرا می شود).

۷ - دربازنمائی[۲۷] هرفریم ۱۰ میلی ثانیه ای قطعات زبانی سیستم TTS، فقط از یک کد ۱۰ بیتی استفاده می نماییم(که متناظر با ایندکس کد آن قطعه در کتاب کد می باشد). البته برای هر نقطه کانتور پیچ و گین، ۲ عدد ۸ بیتی (یکی برای پریود پیچ و دیگری برای اندازه گین) نیز ذخیره می شود (در قسمت LA سیستم TTS).

بنابراین پس از اتمام فرآیند طراحی کتاب کد، تعداد ۱۰۲۴ الگوی زمانی به طول حداکثر ۱۲/۵ میلی ثانیه (و یا ۱۰ میلی ثانیه برای قطعات بی صدا) خواهیم داشت. در بازنمائی قطعات زمانی فقط از ایندکس این الگوهای زمانی استفاده می نماییم.

۳-۳- درصد فشرده سازی روش بازنمایی VQ

فرض می کنیم کوچکترین واحد زبانی مورد استفاده درسیستم TTS، نیمه سیلاب با طول متوسط ۱۵۰ میلی ثانیه باشد (واحد زبانی مورد استفاده در شبیه سازی این تحقیق). تعداد نیمه سیلاب های زبان های زنده در رنج ۲۰۰۰ - ۱۰۰۰ قطعه قرار دارند. فرض می کنیم نمونه های زمانی به صورت ۸ بیتی (با فشرده سازی A-law یا μ-law) ذخیره شده و فرکانس نمونه برداری ۱۱/۰۲۵KHZ باشد (چون برنامه ها تحت سیستم عامل window بوده و این سیستم عامل، فقط سه فرکانس نمونه برداری ۱۱/۰۲۵ و ۲۲/۵ و ۴۴/۱ کیلو هرتز را پشتیبانی می نماید.) اگر قطعات زبانی به صورت معمولی (روش)PSOLA ذخیره شوند، حجم حافظه مورد نیاز برابر خواهد بود با:

$$N_P = 2000 \times 150 \times 11/025 \times 8 = 3/3075 \text{ Mbyte}$$

در روش بازنمایی VQ، دو نوع حافظه مورد نیاز می باشد، الف - حافظه ای که برای ذخیره سازی کد قطعات زمانی لازم است. ب - حافظه ای که برای ذخیره سازی ۱۰۲۴ قطعه زمانی به طول حداکثر ۱۲/۵ میلی ثانیه لازم است. در نتیجه حافظه مورد نیاز روش بازنمائی VQ برابر است با:

$N_V = 2000 \times 15 \times 10 + 1024 \times 12/5 \times 11/025 \times 8 =$ 0.17862 Mbyte

بنابر این نسبت فشرده سازی برابر است با:

$$r = \frac{3/3075}{0/17862} \cong 18/52$$

در نتیجه حجم حافظه مورد نیاز روش بازنمایی VQ حدود ۱۸/۵ برابر کمتر از روش PSOLA گردیده است. تخصیص حافظه به اندازه ۰/۱۸ مگابایت برای قسمت SSسیستم TTS (در سیستم های PC فعلی) امری بسیار مقبول می باشد (در مقابل ۳/۳ مگابایت روش PSOLA).

٤ـ۳ـ روش تغییر پریود پیچ

تمامی قطعات زمانی با صدای ذخیره شده در روشVQ دارای طول ۱۲/۵ms می باشند، در مرحله سنتز نهائی سیستم TTS، علاوه بر تنظیم طول واجگونه ها، ضرورت دارد که کانتور پریود پیچ (وگین) مطابق الگوی مشخص (بسته به نوع لحن و تأکید بیان) تغییر نماید. بنابر این می بایستی طول قطعه را باندازه پریود پیچ متناظر در کانتور پیچ تنظیم نمائیم. این عمل با استفاده از روش های درون یابی (یا چند به یک کردن) مناسب انجام می پذیرد. در این تحقیق از همان روش مطرح در سنتی سایزر PSOLA استفاده شده است.

٤ـ نتایج آزمون های عملی

برای بررسی کیفیت روش ارائه شده در سنتز قابل فهم و طبیعی گفتار، ۲۰ جمله مناسب (تقریباً بالانس از نظر واج های زبان فارسی) طراحی گردید. این جملات توسط دو روش PSOLA و TD-VQSS سنتز گردیدند. از ۱۰ نفر برای مقایسه کیفیت دو روش با همدیگر آزمون های شنیداری انجام پذیرفت. نتیجه آزمون ها در جدول (۲) درج گردیده است. همانطوری که در جدول مشاهده می گردد، کیفیت روش ارائه شده از نظر فهم گفتار و حفظ مشخصات گوینده کاملاً مشابه روش PSOLA است. در مورد طبیعی بودن ۱/۷٪ تنزّل مشاهده می شود. به نظر می رسد که این امر خطای ناشی از تصمیم گیری یا عدم تمرکز شنونده باشد. حال اگر ۱/۷٪ را به عنوان کاهش کیفیت روش TD-VQSS در مقابل روش PSOLA بپذیریم، باتوجه به کاهش ۱۸/۵ برابر در حجم حافظه مورد نیاز (بدون افزایش بار محاسباتی) کاملاً قابل قبول به نظر می رسد.

جدول (۲) مقایسه کیفیت سنتز روش TD-VQSS با روش PSOLA.

	از نظر فهم	از نظر طبیعی بودن	از نظر حفظ مشخصات گوینده
یکسان است	٪۱۰۰	٪۹۸/۳	٪۱۰۰
کمی بدتر است	٪۰	٪۱/۷	٪۰
بدتر است	٪۰	٪۰	٪۰

۵ـ جمع بندی و نتیجه گیری

سنتی سایزر سیگنال گفتار، یک قسمت اساسی از سیستم TTS می باشد. روش هایی که برای سنتز گفتار در این سیستم ها تاکنون مورد استفاده قرار گرفته اند، روش های پارامتریک غیرزمانی می باشند. این روش ها دارای دو اشکال اساسی، الف) عدم تولید صدای کاملاً قابل فهم و طبیعی با حفظ ویژگی های گوینده ب) بالا بودن بار محاسباتی سنتی سایزر در کاربردهای عام منظوره می باشند (بر روی کامپیوترهای PC معمولی و بدون چیپ های مخصوص).

روش موفقی که در مرحله تحقیقاتی مطرح بوده و دو مشکل متذکر را در بر ندارد روش PSOLA می باشد. این روش یک روش سنتز سیگنال گفتار در حوزه زمان است. علت عدم مقبولیت حرفه ای این روش با وجود تولید صدای کاملاً قابل فهم و طبیعی با بار محاسباتی پایین، نیاز به حجم حافظه نسبتاً بالا است. روشی به نام TD-VQSS در طی این تحقیق ارائه گردید که ضمن دارا بودن خواص تقریباً یکسان با روش PSOLA، به حجم حافظه ای حدود ۱۸/۵ برابر کمتر از روش PSOLA نیاز دارد. این مقدار کاهش در حافظه مورد نیاز با استفاده از بازنمایی VQ در حوزه زمان حاصل گردیده است.

زیرنویس ها

1 - time Domain Vector quantization Based Speech Synthesize

2 - Text - to - Speech

3 - Language analyzer

4 - Natural Language Processing Parser

5 - Speech Synthesizer

6 - Text Sentence

7 - Text File
8 - Syllable
9 - Demi Syllable
10 - Diphone
11 - Phone
12 - Concatenation Rules
13 - Pitch Synchronous Overlap Add
14 - Klatt Formant Synthesizer
15 - Line Spectrum Pair Frequency
16 - Epoch Detection
17 - Voiced Frames
18 - Vector Quantization
19 - Interpolation
20 - Decimation
21 - Real Time
22 - Linear Predicting Coding
23 - Line Spectrum Pair Frequency
24 - Line, Bazo, Gray
25 - Frequency Sensitive Learning
26 - Band Limited Interpolation
27 - Representation

مراجع

[1] C. C. Smyth, "Computer Programs of Speech Synthesizing", U.S. Army Human Engineering Lab., Tech. Report, 1989.

[2] I, H. Witten, "Making Computers Talk: An Introduction to Speech Synthesis", Prentice-Hall, Englewood Ciffs, New Jersey, 1986.

[3] J. Martin, "Design of Man-Computer Dialogues", Prentice Hall, Englewood Cliffs, New Jersey, 01973.

[4] H. T. Smith, T. R. G. Green, "Human Interaction with Computers", Academic Press, London, England, 1980.

[5] J. Holmes, "Formant Synthesizers: Cascade or Parallel", Speech Comm. 2, PP. 251-273, 1983.

[6] D. Klatt, "Softare for a Cascade Parallel Formant Synthesizer", Speech Comm. 2, PP. 251-273, 1980.

[7] C. Browman, "Rules for Demisyllabe Synthesis Using LINGWA, a Langue Interpreter", Proc. IEEE on ICASSP, PP. 561-592, 1980.

[8] H. Dettweiler, W. Hess, " Concatenation Rules for Demisyllable Speech Synthesis", Acustica, 57, PP. 268-283, 1985.

[9] D. Klatt, "Text to Speech: Present and Future", Speech Tech., PP. 221-226, 1986.

[10] C. Hamon, F. Charpentier," A Diphone Synthesis System Based on Time Domain Prosodic Modification of Speech", IEEE Proc. on ICASSP, 1989.

[11] D. Bigorgne, S. White, "Multilingual PSOLA Text to Speech System", IEEE Proc. on IGASSP, 1993.

[12] H. Kwa, S. Yamamoto, "Development of a Text to Speech for Japanese Based on Waveform Splicing", IEEE Proc. on ICASSP, 1994.

[13] F. M. Galanes, m. H. Savoji, H. M. Pardo, "Speech Synthesis System Based on a Variable Decimation Interpolation Factor", IEEE Proc. on ICASSP 1995.

[14] J. Mullen, "Speech Synthesis IC for Low-Cost Application", New Electronic, PP. 99-101, June, 1994 .

[15] G. Kaplan, E. Lerner, "Realism in Synthetic Speech", IEEE Spectrum, 22, PP. 32-37, April 1994.

[16] R. W. Schafer, J. Markel, "Speech Analysis", IEEE Press 1979.

[17] R. M. Gray, "Vector Quantization", IEEE ASSP Magazine, April 1982.

[18] S. C. Ahalt, P. Chen, "Vector Quantization Using Frequency SensitiveLearning", IEEE Proc. on ICSE, 1989.

[19] A. V. Opeenhein, "Advance Topics in Signal Processing", Prentice Hall, 1988.

[20] P.P. Vaidyanathan, "Multitate Systems and Filter Baaks", Prentice Hall, 1993.